Actualidades en Psicología, 31(122), 2017, 133-148 ISSN 2215-3535
http://revistas.ucr.ac.cr/index.php/actualidades DOI: http://dx.doi.org/10.15517/ap.v31i122.23499

El Modelo de Crédito Parcial aplicado a la escala Distorsión del Big Five Questionnaire

The Partial Credit Model applied to the Lie Scale of the Big Five Questionnaire

Facundo Juan Pablo Abal[1]
Gabriela Susana Lozzia[2]
Sofía Esmeralda Auné[3]
Horacio Félix Attorresi[4]
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Universidad de Buenos Aires, Argentina

Resumen. Se aplicó el Modelo de Crédito Parcial (MCP) de la Teoría de Respuesta al Ítem al análisis de ítems de la adaptación española escala Distorsión del Big Five Questionnaire. Esta escala evalúa la tendencia de los individuos a ofrecer un perfil distorsionado. Participaron 1592 adultos de población general (55% sexo femenino). El análisis de los datos se realizó con Winsteps. El ajuste del MCP fue adecuado para todos los ítems; no obstante, un porcentaje considerable de evaluados no presentó un patrón de respuestas acorde a las expectativas del modelo. Cinco ítems presentaron inversiones en el orden esperable para los valores estimados de parámetros de umbral. Los resultados revelaron las debilidades que presenta la escala y orientan sobre posibles modificaciones futuras.

Palabras clave. Modelo de Crédito Parcial, Escala Distorsión, deseabilidad social, Cuestionario Big Five, Teoría de Respuesta al Ítem.

Abstract. The Partial Credit Model (PCM) of the Item Response Theory was applied to the Spanish Lie Scale adaptation of the Big Five Questionnaire. The scale measures individuals' tendency to provide a distorted profile. The sample comprised 1592 adults from the general population (55% females). All analyses were performed by means of Winsteps software. The PCM exhibited satisfactory goodness-of-fit for all items. However, a considerable proportion of respondents had incongruent response patterns which were not in agreement with the model's expectations. Five items presented inversions in the order expected for the estimated values of threshold parameters. These findings show the scale weaknesses and yield useful information to guide possible changes in future research.

Keywords. Partial Credit Model, Lie Scale, Social Desirability, Big Five Questionnaire, Item Response Theory.

[1]Facundo Juan Pablo Abal. Universidad de Buenos Aires, Argentina.Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Dirección postal: Zuviría 5691, C.P.: 1439, Ciudad de Buenos Aires. E-mail: fabal@psi.uba.ar
[2]Gabriela Susana Lozzia.Universidad de Buenos Aires, Argentina. E-mail: glozzia@psi.uba.ar
[3]Sofía Esmeralda Auné.Universidad de Buenos Aires, Argentina. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). E-mail: sofiaaune177@hotmail.com
[4]Horacio Félix Attorresi.Universidad de Buenos Aires, Argentina. E-mail: horacioattorresi@gmail.com

Introducción

Las distorsiones de respuesta son una de las amenazas más importantes que tiene la evaluación mediante tests de comportamiento típico. Ya sea un intento deliberado por parte del evaluado de falsear sus respuestas (simulación) o una tendencia involuntaria a atribuirse a sí mismo cualidades socialmente deseables (deseabilidad social), las distorsiones constituyen un factor ajeno e independiente al atributo psicológico medido que influye en el resultado final de la medición.

El *Big Five Questionnaire* es un instrumento psicométrico que mide características de la personalidad basado en el Modelo de los Cinco Factores e incluye entre sus ítems una escala que permite registrar esta tendencia a la distorsión en las respuestas (Caprara, Barbaranelli & Borgogni, 1993). La escala *Lie*, denominada escala Distorsión en la versión española de Bermúdez (1995), tiene como objetivo la detección de la tendencia de un individuo a ofrecer un perfil distorsionado, ya sea en un sentido positivo (*faking good*) o negativo (*faking bad*). Los ítems reflejan conductas cotidianas estimadas socialmente como buenas o malas que resultan poco probables de ser realizadas concretamente de manera tan frecuente como propone el enunciado. De esta manera, se asume que la respuesta de los individuos puede estar influida por la deseabilidad social o por una intención deliberada de brindar una buena imagen en la medida en que estas respuestas tienen baja probabilidad de ser realizadas. Las frases incluyen adverbios de tiempo tales como siempre o nunca a fin de garantizar que las respuestas más extremas sean improbables.

La escala Lie ha presentado propiedades psicométricas aceptables, tanto en su versión original en italiano (Caprara et al., 1993), como en las diferentes adaptaciones que ha tenido el instrumento al español, inglés y alemán (Barbaranelli & Caprara, 2002; Caprara, Barbaranelli, Bermúdez, Maslach & Ruch, 2000; Caprara, Barbaranelli & Borgogni, 1996). El factor común de estos estudios de validación es que fueron realizados en el marco de la Teoría Clásica de los Tests (TCT) y que, por ende, presentan las limitaciones propias de este modelo psicométrico. Desde hace ya varios años la Psicometría mundial se encuentra atravesando un período de transición hacia la Teoría de la Respuesta al Ítem (TRI) provocando una revolución en la forma de construir y de administrar los tests. Una de las principales ventajas que presenta la TRI por sobre la TCT es que permite realizar un análisis de los ítems más profundo y exhaustivo. Como consecuencia, puede revelar problemas que quedaban inadvertidos para los criterios de calidad definidos en la TCT (Martínez Arias, 1995).

La primera generación de los modelos construidos en el marco de la TRI fue desarrollada en un contexto educativo para pruebas de ejecución máxima que solo permitían analizar ítems puntuados de forma dicotómica. Sin embargo, la TRI ha tenido una progresiva evolución, generando nuevos modelos para el análisis de ítems politómicos y también extendiendo su aplicación a constructos evaluados mediante tests de comportamiento típico. En esta línea, el Modelo de Crédito Parcial (MCP) de Masters (1982, 2016) ha sido utilizado con frecuencia para la modelización de actitudes y rasgos de la personalidad (e.g. Abal, Galibert, Aguerri & Attorresi, 2014; DiStefano, Morgan & Motl, 2012; Rojas & Pérez, 2001; Vendramini, Silva & Dias, 2009).

El MCP es una extensión del modelo desarrollado por Rasch (1960) para ítems dicotómicos. Rasch definió la relación entre el comportamiento de un individuo y la probabilidad $P_i(1 \mid \theta)$ de que este obtenga la puntuación 1 en el ítem dicotómico *i* mediante una función logística:

$$P_i\left(1|\theta\right) = \frac{e^{(\theta-\beta_i)}}{1+e^{(\theta-\beta_i)}}$$

Esta probabilidad depende del nivel del rasgo latente θ del individuo y de la dificultad β_i del ítem. Masters y Wright (1997) utilizaron la denominación paso para describir la transición de la categoría 0 a la categoría 1. El Modelo de Rasch (MR), por ser dicotómico, tiene un solo paso y el parámetro βi del ítem *i* especifica la cantidad de rasgo necesario para efectuar esta transición; en el sentido de tener más chance de elegir

la categoría 1 en lugar de elegir la 0. Esta transición depende de si el individuo cuenta con un nivel de θ suficiente como para superar el parámetro β_i. En ítems de tests de comportamiento típico, el parámetro β_i se interpreta como un punto de transición (dentro de la escala del rasgo) entre la probabilidad de tomar al enunciado como no-descriptivo del evaluado y la de considerarlo como descriptivo (Panter, Swygert & Dahlstrom, 1997).

Wright y Stone (1979) realizaron una descripción del MR que es útil para comprender su generalización a ítems politómicos. Estos autores definieron que la probabilidad de que la persona obtenga una puntuación de 1 en el ítem i puede expresarse como:

$$\frac{P_i\left(1|\theta\right)}{P_i\left(0|\theta\right)+P_i\left(1|\theta\right)}=\frac{e^{(\theta-\beta_i)}}{1+e^{(\theta-\beta_i)}}$$

Como los ítems modelizados con el MR solo presentan dos categorías, la probabilidad $P_i(0 \mid \theta)$ de puntuar 0, sumada a la probabilidad $P_i(1 \mid \theta)$ de obtener la puntuación 1, debe dar como resultado 1 para todo nivel del rasgo latente y, en consecuencia, esta formulación es equivalente a la desarrollada previamente. Siguiendo esta expresión, Masters (1982) pudo realizar la extensión del MR a ítems politómicos con formato de respuesta ordenada de m+1 categorías sin alterar sustancialmente la formulación original. El MCP permite calcular la probabilidad que tiene un individuo con rasgo θ de elegir la categoría h ($h = 0,\ldots, m$) en el ítem i a partir de una segmentación adyacente del dato politómico. De esta manera, es posible definir un modelo general para toda categoría h con $h = 0,\ldots, m$ como:

$$\frac{P_i\left(h|\theta\right)}{P_i\left(h-1|\theta\right)+P_i\left(h|\theta\right)}=\frac{e^{(\theta-\beta_{ih})}}{1+e^{(\theta-\beta_{ih})}} \qquad (1)$$

$P_i(h \mid \theta)$ es la probabilidad de que un sujeto de rasgo θ opte por la categoría h en el ítem i. $P_i(h\text{-}1 \mid \theta)$ es la probabilidad de que una persona de rasgo θ elija la categoría anterior a h (o sea h-1) en el ítem i. El parámetro β_{ih} es un umbral que separa la transición entre la categoría h-1 y h en el ítem i. El MCP se focaliza en la cantidad de rasgo que demanda el ítem para que el evaluado tenga más chances de elegir la categoría h en lugar de la inmediatamente anterior ($h - 1$). Para un valor de β_{ih}, el individuo tiene la misma probabilidad de responder en la categoría h o en $h - 1$. Lógicamente, este parámetro está definido solamente para $h = 1,\ldots,$ m porque no existe una categoría anterior a $h = 0$ y, por ende, tampoco existe el umbral β_{i0}.

A diferencia del MR, que solo presenta un parámetro β_i, el MCP tiene una serie de valores de umbral β_{ih} que definen, en la escala del rasgo, las localizaciones de cada uno de los pasos que contiene el ítem. En ítems que miden rendimiento o habilidad, los pasos representan el conocimiento parcial del examinado; ya sea porque se aceptan respuestas parcialmente correctas, porque reciben calificaciones graduadas en distintos niveles de resolución o ambas situaciones. Para los ítems de tests de comportamiento típico no existe un registro tangible de las transiciones que el evaluado realiza de una categoría a otra. Simplemente debe suponerse, a partir de la respuesta observada en la escala Likert, que la persona ha dado los pasos previos. Conviene aclarar que los pasos de los ítems politómicos están secuenciados. De esto se desprende que solo cuando un individuo disponga de un nivel de rasgo suficiente para realizar la transición de una categoría (h-1) a la siguiente h, podrá enfrentarse al próximo paso.

$P_i(h \mid \theta)$ se presenta en forma implícita en la ecuación (1) pero se necesita explicitarla para la construcción del Modelo de Crédito Parcial. Partiendo de (1) se tiene:

$$P_i\left(h|\theta\right)\left[1+e^{(\theta-\beta_{ih})}\right]=\left[P_i\left(h-1|\theta\right)+P_i\left(h|\theta\right)\right]e^{(\theta-\beta_{ih})}$$

Aplicando la propiedad distributiva en ambos términos se obtiene:

$$P_i\left(h|\theta\right)+P_i\left(h|\theta\right)e^{(\theta-\beta_{ih})}=P_i\left(h-1|\theta\right)e^{(\theta-\beta_{ih})}+P_i\left(h|\theta\right)e^{(\theta-\beta_{ih})}$$

Resultando la expresión:

$$P_i\left(h\middle|\theta\right) = P_i\left(h-1\middle|\theta\right)e^{(\theta-\beta_{ih})} \quad \text{para h = 1,..., } m$$

Esto implica que:

$$P_i\left(1\middle|\theta\right) = P_i\left(0\middle|\theta\right)e^{(\theta-\beta_{i1})}$$
$$P_i\left(2\middle|\theta\right) \quad P_i\left(1\middle|\theta\right)e^{(\theta \ \beta \ )}$$
$$\text{----------}$$
$$P_i\left(m\middle|\theta\right) = P_i\left(m-1\middle|\theta\right)e^{(\theta-\beta_{im})}$$

Dado que las fórmulas correspondientes a cada una de las categorías se encuentran encadenadas, es posible expresarlas de manera general considerando la regla del producto de exponentes con igual base:

$$P_i\left(h\middle|\theta\right) = P_i\left(0\middle|\theta\right)e^{\sum_{k=1}^{h}(\theta-\beta_{ik})} \quad \text{para} \quad h = 0,..., m \quad (2)$$

Por último, falta calcular $P_i(0 \mid \theta)$. Como es:

$$P_i(0\middle|\theta) + P_i\left(1\middle|\theta\right) + P_i\left(2\middle|\theta\right) + ... + P_i\left(m\middle|\theta\right) = 1$$

Entonces:

$$P_i(0\middle|\theta) + e^{(\theta-\beta_{i1})}P_i\left(0\middle|\theta\right) + e^{\left[(\theta-\beta_{i1})+(\theta-\beta_{i2})\right]}P_i\left(0\middle|\theta\right) + ...$$
$$+e^{\left[(\theta-\beta_{i1})+(\theta-\beta_{i2})+...+(\theta-\beta_{im})\right]}P_i\left(0\middle|\theta\right) = 1$$

O lo que es lo mismo

$$P_i(0\middle|\theta) + e^{\sum_{h=1}^{1}(\theta-\beta_{ih})}P_i\left(0\middle|\theta\right) + e^{\sum_{h=1}^{2}(\theta-\beta_{ih})}P_i\left(0\middle|\theta\right) + ... + e^{\sum_{h=1}^{m}(\theta-\beta_{ih})}P_i\left(0\middle|\theta\right) = 1$$

Sacando como factor común $P_i(0 \mid \theta)$ se obtiene que:

$$P_i(0\middle|\theta)\left[1 + e^{\sum_{h=1}^{1}(\theta-\beta_{ih})} + e^{\sum_{h=1}^{2}(\theta-\beta_{ih})} + ... + e^{\sum_{h=1}^{m}(\theta-\beta_{ih})}\right] = 1$$

$$P_i(0\middle|\theta) = \frac{1}{1 + e^{\sum_{h=1}^{1}(\theta-\beta_{ih})} + e^{\sum_{h=1}^{2}(\theta-\beta_{ih})} + ... + e^{\sum_{h=1}^{m}(\theta-\beta_{ih})}}$$

Con el fin de hacer más compacta la fórmula, se reemplaza el 1 ubicado en el denominador definiendo:

$$\theta - \beta_{i0} = \sum_{h=0}^{0}\left(\theta - \beta_{ih}\right) = 0$$

Es importante destacar que esto se realiza por definición y con el único fin de simplificar la fórmula porque, según la formulación del modelo, el parámetro β_{i0} no existe. A partir de esto,

$$1 = e^{0} = e^{\sum_{h=0}^{0}(\theta-\beta_{ih})}$$

Por consiguiente es posible expresar $P_i(0 \mid \theta)$ como:

$$P_i(0\middle|\theta) = \frac{1}{e^{\sum_{h=0}^{0}(\theta-\beta_{ih})} + e^{\sum_{h=0}^{1}(\theta-\beta_{ih})} + e^{\sum_{h=0}^{2}(\theta-\beta_{ih})} + ... + e^{\sum_{h=0}^{m}(\theta-\beta_{ih})}}$$

Lo que puede expresarse como:

$$P_i(0\middle|\theta) = \frac{1}{\sum_{j=0}^{m} e^{\sum_{h=0}^{j}(\theta-\beta_{ih})}}$$

Reemplazando $P_i(0 \mid \theta)$ en (2), se concluye:

$$P_i(h\middle|\theta) = \frac{e^{\sum_{k=0}^{h}\theta-\beta_{ik}}}{\sum_{j=0}^{m} e^{\sum_{h=0}^{j}\theta-\beta_{ih}}} \text{ , para } h=0,..., m \text{ y } \beta_{ih} \text{ con k=1,..., } m.$$

Donde se define: $\sum_{h=0}^{0}\left(\theta - \beta_{ih}\right) = 0$

La expresión obtenida muestra cómo la probabilidad de responder a una categoría h se representa como una exponencial correspondiente a esta categoría, que se divide por la sumatoria de las exponenciales de todas las categorías (Embretson & Reise, 2000). Básicamente, el modelo deduce que la probabilidad que tiene una persona de escoger *h* se corresponde con la diferencia entre su nivel de rasgo θ y el parámetro de umbral β_h asociado a cada categoría, contemplando todos los pasos del ítem en conjunto.

Objetivo

Ha pasado más de medio siglo desde que Lord (1952) y Rasch (1960) publicaron sus primeras ideas sobre la TRI y, sin embargo, su aplicación a tests de comportamiento típico es relativamente escasa (Morizot, Ainsworth & Reise, 2007; Reise & Revicki, 2015 ; Reise & Waller, 2009, Thomas, 2011). Parte de esta limitada implementación se explica por la poca difusión de las ventajas que presentan los modelos de la TRI al ser usados en el análisis de los ítems de personalidad y actitudes (Abal, Lozzia, Aguerri, Galibert & Attorresi, 2010). En particular, desde la TRI se ha mostrado un exiguo desarrollo en la modelización de constructos vinculados con la distorsión de las respuestas a los cuestionarios. La mayoría de las investigaciones en esta área se han abocado a la aplicación de modelos dicotómicos (Ferrando & Chico, 2000; Seol, 2007; Vésteinsdóttir, Reips, Joinson & Thorsdottir, 2017) a la escala de Deseabilidad Social de Marlowe y Crowne (Crowne & Marlowe, 1960). También, se ha registrado el uso de modelos politómicos (Asgeirsdottir, Vésteinsdóttir & Thorsdottir, 2016; Vispoel & Kim, 2014) en el Inventario Balanceado de Respuesta Deseable de Paulhus (1988).

A la luz de estas consideraciones teóricas y metodológicas, se propone como objetivo de este trabajo estudiar la calidad psicométrica de los ítems de la escala Distorsión del Big Five Questionnaire (Caprara et al., 1993, adaptación de Bermúdez, 1995) a partir de su modelización con el Modelo de Crédito Parcial de Masters (1982, 2016). En este sentido, en virtud de que la escala ha sido válida solo desde la perspectiva clásica, se pretende identificar potenciales problemas no detectados previamente en el test gracias a la exhaustividad en el diagnóstico de los ítems que brinda la aplicación de un modelo politómico de la TRI.

Método

Participantes

Se contó con la colaboración de una muestra de 1592 personas adultas residentes en Ciudad de Buenos Aires (Argentina). El 55% fueron mujeres y el 45% fueron varones. La edad de los participantes osciló entre 18 y 74 años con un promedio de 35.8 años (*DT* = 13.2). Las ocupaciones más frecuentemente reportadas fueron: empleado (40.6%), docente (11.1%) y estudiante (10.9%). Con respecto al nivel educativo, el 37.2% alcanzó como máximo estudios secundarios completos, el 19.6% poseía o estaba cursando estudios terciarios y el resto refirió tener estudios universitarios incompletos (26.9%) y completos (16.3%). Se trabajó con un muestreo no probabilístico por accesibilidad. Para la determinación del tamaño de la muestra se recurrió al requerimiento explicitado por Linacre (1994) que solicita al menos 10 sujetos por categoría de respuesta para alcanzar una adecuada estimación de los parámetros del MCP. Los participantes firmaron un consentimiento informado en el que se detallaban los objetivos del estudio, así como también se brindaban las garantías de protección de sus datos personales y la confidencialidad de sus respuestas.

Instrumento

*Escala de Distorsión del Big Five Questionnaire (*Caprara et al., 1993, adaptación de Bermúdez, 1995). La prueba mide la deseabilidad social, entendida como un estilo de respuesta del sujeto en el que busca mostrar aspectos favorables de su personalidad y eludir los menos favorables. Consta de 12 ítems con cinco opciones de respuesta (de *Completamente falso para mí a Completamente verdadero para mi*). La escala fue diseñada considerando que los puntajes elevados son indicativos de una tendencia a proporcionar una respuesta más o menos intencionada que refleja una imagen de sí mismo artificiosamente positiva. Por el contrario, puntajes bajos en extremo describen a los evaluados que tendieron a dar una imagen negativa de sí mismos. Todos están redactados en forma positiva, lo cual implica que las respuestas que afirman la veracidad de la sentencia (*Bastante verdadero y Completamente verdadero*) se asocian a cualidades socialmente deseables mientas que Bastante falso y Completamente falso describen comportamiento sociales indeseables.

En el estudio factorial efectuado para la adaptación española se identificaron dos dimensiones para la escala Distorsión (Bermúdez, 1995). En la primera,

los ítems apuntaban a recoger información sobre atributos deseables relacionados con el yo (e.g. ser eficaz para resolver problemas, valiente o seguro de sí mismo); mientras que, en la segunda, se describían comportamientos deseables en la interacción con otros (e.g. ser obediente, sincero o cordial). No obstante, ni los baremos ni los estudios de confiabilidad recogen estos aspectos diferenciales del constructo. Asimismo, estas subdimensiones tampoco aparecen reportadas por los autores de la versión original (Caprara et al., 1993). En relación con la confiabilidad de la escala, Bermúdez (1995) refiere en el manual de la adaptación del BFQ un índice alfa de Cronbach de .77. Para la presente investigación se halló un índice ligeramente más bajo de .74. El estudio de consistencia interna mediante indicadores alternativos al alfa de Cronbach presentó mejores resultados (Alfa ordinal = .82 y *Greatest Lower Bound* = .87).

Procedimiento

Se adoptó un procedimiento acorde con las características de un estudio instrumental (Ato, López & Benavente, 2013; Carretero-Dios & Pérez, 2007) tendiente a revisar las evidencias de validez de la escala basadas en la estructura interna. En esta línea, la TRI ofrece la posibilidad de examinar la relación entre las categorías de respuesta a cada ítem y la variable latente que el indicador pretende medir.

A pesar de que fueron utilizados ítems redactados en idioma español, se consideró pertinente realizar un análisis de contenido de ellos previa administración del instrumento. Los ítems fueron revisados por cinco jueces locales expertos en evaluación psicológica y psicometría, con el objetivo de dar garantías de la adecuación a los giros idiomáticos argentinos y la equivalencia conceptual del instrumento. La conclusión obtenida en este estudio mostró que no era necesario efectuar modificaciones.

Los participantes fueron informados de manera oral y escrita sobre las condiciones en las que se llevaría a cabo el estudio y la futura utilización de los datos con fines investigativos. Se les explicó que la prueba tenía como objetivo la medición de atributos de su personalidad y que no había respuestas a los ítems que pudieran ser consideradas como correctas o incorrectas. Asimismo, se enfatizó sobre las garantías del anonimato y confidencialidad, con el fin de propiciar una baja motivación para el disimulo.

Los examinadores fueron debidamente entrenados y tomaron los datos individualmente en entornos acordes con las coordenadas deseables para un adecuado *setting* de evaluación. El protocolo administrado incluyó otras escalas no consideradas para los fines de esta investigación. Los individuos respondieron el cuestionario de forma autoadministrada y sin tiempo límite. En todo momento se garantizó el bienestar de los examinados, a quienes se les advirtió sobre la posibilidad de cesar con su colaboración en cualquier punto de la evaluación. Los sujetos no recibieron ningún tipo de compensación por su participación.

Análisis de datos

Se verificó el supuesto de unidimensionalidad del constructo requerido por el MCP operando con el software Mplus (Muthén & Muthén, 2010). Se aplicó un Análisis Factorial Confirmatorio (AFC) respetando el carácter ordinal de los datos. En consecuencia, la estimación de los parámetros se realizó con el método de mínimos cuadrados ponderados robustos (*Weighted Least Squares Mean and Variance Adjusted*, WLSMV) sobre la base de las matrices de correlaciones policóricas. El ajuste al modelo unidimensional se analizó considerando los índices CFI (*Comparative Fit Index*), TLI (*Tucker-Lewis Index*) y RMSEA (*Root Mean Square Error of Approximation*).

La aplicación del MCP se realizó con el programa Winsteps versión 3.63.0 (Linacre, 2006). La estimación de los parámetros se efectuó por el Método de Máxima Verosimilitud Conjunta. En virtud de la cantidad de categorías que presenta la escala Likert del instrumento modelizado, se estimaron 48 parámetros de los ítems (β_1, β_2, β_3 y β_4 para cada uno de los 12 ítems) y 1592 parámetros θ que representan el nivel de rasgo de los individuos participantes. El ajuste del MCP a los datos se estudió a nivel global y también ítem a ítem a partir de los valores de los estadísticos ajuste próximo

(Infit) y lejano (Outfit). El Infit permite detectar la presencia de patrones de respuesta anómalos (i.e. comportamientos no esperados según las predicciones del modelo) en ítems cuya zona de actuación se encuentra cercana al nivel de rasgo del sujeto. En cambio, el Outfit resulta sensible a comportamientos extremos no esperados que se encuentren alejados al nivel de rasgo del evaluado. Los Infit y Outfit se interpretan como medias cuadráticas de los residuales (Mean-Square, MnSq). Ambos indicadores tienen una expectativa de 1, valor que adoptan frente a un ajuste perfecto entre los datos y el modelo. Si bien se espera que estos resultados sean alcanzados en la evaluación del ajuste global, el ajuste pormenorizado de los ítems establece una zona de ajuste aceptable en el intervalo de 0.5 a 1.5 para ambas MnSq (Linacre, 2012). Winsteps también ofrece las medias cuadráticas de los residuos estandarizados (ZSTD), pero este indicador no fue considerado porque resulta sensible para un tamaño de muestra elevado como el que tiene la presente investigación (Linacre, 2012).

Resultados

Supuesto de unidimensionalidad

Tanto los indicadores de ajuste comparativo CFI = .93 y TLI = .91 como el índice de ajuste absoluto RMSEA=.074, IC 90% [.068, .079] adoptaron valores aceptables. Aun cuando no existe un consenso acabado sobre los puntos de cortes de estos índices de ajuste (Abad, Olea, Ponsoda & García, 2011; Brown, 2015; Byrne, 2012), es posible aceptar razonablemente el cumplimiento del supuesto de unidimensionalidad dado que CFI y TLI fueron mayores a .90 y RMSEA fue inferior a .08. Considerando los resultados obtenidos en la adaptación española, se estudió el ajuste a un modelo que contempla las dos subdimensiones identificadas por Bermúdez (1995). Si bien los indicadores de ajuste mejoraron levemente, CFI = .94, TLI = .92, RMSEA=.067, IC 90% [.061, .073], la correlación interfactor resultó elevada (r = .85) según el criterio establecido por Kline (2011). Por esta razón, resulta más parsimonioso sostener el modelo que establece una única dimensión.

Calibración de los ítems

El programa requirió realizar 27 iteraciones para que la estimación alcance el criterio de convergencia de .00001, lo cual resulta una cantidad de ciclos razonable. Los parámetros presentaron valores dentro de un rango esperable y los errores de estimación resultaron relativamente bajos. En la tabla 1 se presentan los indicadores que permiten estudiar el ajuste del MCP a los datos de manera global para las personas y los ítems.

Si se observan los resultados referidos a los sujetos es posible apreciar que los valores de θ se distribuyen con media -0.94 y desviación estándar de 0.74. Los

Tabla 1

Ajuste global del Modelo de Crédito Parcial.

	Ajuste global de sujetos				Ajuste global de Ítems			
	θ	*Se*	Infit MnSq	Outfit MnSq	Nivel de adhesión	*Se*	Infit MnSq	Outfit MnSq
Media	-0.94	0.35	1.04	1.02	0.00	0.03	1.00	1.02
DT	0.74	0.10	0.61	0.69	0.44	0.00	0.10	0.14
Máx	1.12	1.01	4.55	9.66	0.63	0.03	1.26	1.30
Mín	-3.72	0.27	0.07	0.08	-0.55	0.02	0.86	0.80

Nota. θ = Nivel de la variable; Se: Error de estimación; MnSq = Media cuadrática de los residuales del ajuste interno (Infit) y ajuste externo (Outfit).

promedios de las MnSq han sido próximos a 1 para los Infit y Outfit pudiendo concluir que el modelo ajusta razonablemente a los datos. No obstante, se observan valores máximos y mínimos que se ubican por fuera del rango de ajuste aceptable. Esto implica que existe un porcentaje de sujetos cuyos patrones de respuesta presentan desajustes y, por ende, no pueden ser explicados por el MCP. Es importante destacar que se trata de un 18% y 15.7% de personas según se tome el Infit y Outfit, respectivamente.

La evaluación global del ajuste de los ítems también permite corroborar que el modelo tuvo un ajuste satisfactorio a los datos en virtud de que se registraron promedios cercanos a 1 para las MnSq de Infit y Outfit. En este caso, todos los ítems presentaron Infit y Outfit dentro de los límites aceptables. La tabla 1 muestra además el nivel de adhesión promedio de los ítems (medido en la misma escala del rasgo) que es fijado a 0 en el proceso de estimación que realiza Winsteps. Este nivel de adhesión medio surge de promediar los niveles de adhesión de cada uno de los ítems (tabla 2). A su vez, este se obtiene de promediar los parámetros β_h que separan cada una de las categorías de un mismo reactivo. De esta manera, el nivel de adhesión del ítem permite caracterizar su ubicación en términos generales, delimitando una zona de actuación donde sus opciones de respuesta tienen mayor probabilidad de elección.

Si se compara el promedio 0 de nivel de adhesión de los ítems con el valor promedio de θ = -0.94, se puede concluir que los evaluados tendieron a concentrarse en los niveles más bajos del rasgo con un θ mínimo de -3.72 y un máximo de 1.12. En efecto, el 91.8% de los participantes adoptó valores negativos de θ reflejando de una manera masiva una imagen de sí mismos escasamente influenciada por la deseabilidad social y con baja predisposición al falseamiento. Incluso este resultado parece mostrar una mayor tendencia de los participantes a ofrecer una imagen negativa de sí mismos.

Este análisis es posible debido a que las distribuciones de los niveles de adhesión de los ítems y los niveles del rasgo de los individuos se encuentran en la misma métrica. La figura 1 ofrece una visión general del rango de actuación de los ítems en relación con las medidas de las personas evaluadas. Conviene señalar que la alta concentración en torno a 0 de los niveles de adhesión de los ítems se debe a que en la figura 1 se representan los promedios de los parámetros β_h para cada ítem. Al considerar los β_h más extremos se podría definir un rango de actuación de los ítems entre -2.30 (β_1 del ítem 3) y 1.90 (β_4 del ítem 1).

En la tabla 2 también se incluyen los indicadores de ajuste (Infit y Outfit) y los parámetros de umbral correspondientes a cada reactivo. Aunque los parámetros β_{ih} estimados para los ítems de la escala aparecen en su mayoría ordenados de forma creciente, se puede apreciar que la modelización de cinco de ellos (ítems 2, 4, 6, 8 y 9) registra inversiones en la secuencia (*reversals*). La forma en que deben interpretarse estas inversiones es objeto de una fuerte controversia en la actualidad (Adams, Wu & Wilson, 2012; Andrich, 2013; García-Pérez, 2017; Wetzel & Carstensen, 2014). La lectura más aceptada de estos resultados supone que la inversión en orden de los parámetros βh de un ítem conlleva una violación del orden que se le supone a las categorías de respuesta de la escala Likert (Andrich, 2013; Embretson & Reise, 2000). A modo de ejemplo, en el ítem 2, el último parámetro de umbral (β_4 = 0.76) presenta una localización en la escala del rasgo que es inferior al anteúltimo parámetro del ítem (β_3 = 1.89). Según el planteo del MCP, una persona A cuyo nivel de rasgo θ es de 0.76 (o sea que $\theta_A = \beta_4$) tendría idéntica probabilidad de elegir las categorías *Bastante verdadero para mí* o *Completamente verdadero para mí.* Sin embargo, otra persona B cuyo nivel de Distorsión es θ_B = 1.89 tendría la misma probabilidad de escoger N*i verdadero ni falso para mí* o *Bastante verdadero para mí.* Esto significa que A, aunque tenga un menor nivel de rasgo que B ($\theta_A < \theta_B$) podría optar por una categoría en la escala Likert que representa a un mayor nivel en la variable.

Wetzel y Carstensen (2014) sugirieron profundizar el estudio de la relación entre la escala del rasgo y la escala Likert, a partir del θ promedio obtenido por los evaluados que escogieron cada una de las categorías. En la tabla 3 se puede observar que estos promedios

Tabla 2

Estimación y ajuste de los ítems

Ítem	*Nivel de adhesión*	*Se*	Infit	Outfit	Parámetros de umbral			
			MnSq	MnSq	β_1	β_2	β_3	β_4
1	0.45	0.03	0.96	1.00	-0.07	-0.05	0.03	1.90
2	0.46	0.03	0.97	1.01	-0.58	-0.23	1.89	0.76
3	-0.55	0.03	0.99	0.99	-2.30	-1.50	-0.11	1.69
4	0.20	0.03	0.99	1.09	-0.52	0.47	0.06	0.79
5	-0.53	0.03	0.97	0.97	-1.78	-1.08	-0.26	1.00
6	0.01	0.03	0.93	0.93	-0.69	-0.10	-0.13	0.98
7	-0.47	0.02	1.26	1.30	-1.30	-0.83	-0.22	0.47
8	0.63	0.03	0.89	0.81	0.50	0.85	-0.01	1.19
9	-0.14	0.03	1.09	1.14	-1.48	-0.59	0.77	0.75
10	-0.22	0.03	1.04	1.05	-1.75	-0.71	-0.16	1.73
11	0.61	0.03	0.86	0.80	-0.25	0.39	1.07	1.22
12	-0.46	0.02	1.11	1.16	-1.27	-0.89	-0.47	0.79

Nota. Se = Error de estimación; MnSq = Media cuadrática de los residuales del ajuste interno (Infit) y ajuste externo (Outfit).

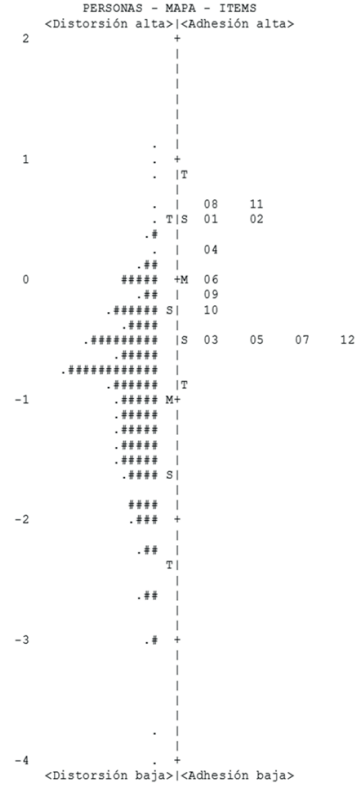

Figura 1. Mapa de personas-ítems. Cada "#" = 15 personas. Cada "·" = entre 1 y 14 personas. M = Media en la distribución de personas o ítems. S = 1 desvío a la media en las distribución de las personas o ítems. T = 2 desvíos a la media en las distribución de las personas o ítems.

no se incrementaron monótonamente en los ítems 8 y 9. Por ejemplo, quienes contestaron *Completamente verdadero para mí* en el ítem 8 tuvieron, en promedio, un θ = -0.25, valor que se encuentra por debajo del promedio alcanzado por quienes escogieron *Bastante verdadero para mí* (θ = -0.09) y *Ni verdadero ni falso para mí* (θ = -0.1).

Análisis de las Curvas Características

En la figura 2 aparecen las Curvas Características de las Categorías de Respuesta, que permiten describir la probabilidad que tiene una persona de elegir cada una de las categorías de la escala Likert en función de su nivel de rasgo latente. En el eje de ordenadas se representa la probabilidad de respuesta a cada una de las cinco categorías del ítem, en tanto que en el eje de abscisas se distribuye el rasgo latente. Otra interpretación que se desprende del mismo gráfico es la posibilidad de establecer el porcentaje esperado de personas que optarán por cada una de las opciones del ítem en función del nivel de Distorsión que manifiesten en sus respuestas.

El análisis de estas curvas brinda información sobre las zonas del rasgo latente en que cada una de las categorías es más probable. A su vez, permite verificar si todas las opciones alcanzan a ser máximamente probables en algún intervalo del recorrido del rasgo. Esto constituye una importante propiedad psicométrica de los ítems de la escala en tanto que garantiza que todas las categorías de la escala Likert son útiles para discriminar en algún rango específico de la variable. La inspección de las curvas permite apreciar claramente cuál es la consecuencia de la existencia de inversiones

Tabla 3

Medias y desvíos estándar de las θ de las personas que escogieron cada categoría.

Ítem	Completamente falso para mí	Bastante falso para mí	Ni verdadero ni falso para mí	Bastante verdadero para mí	Completamente verdadero para mí
1	-1.30 (0.03)	-0.77 (0.03)	-0.35 (0.03)	-0.29 (0.04)	-0.28 (0.18)
2	-1.37 (0.03)	-0.75 (0.02)	-0.43 (0.03)	-0.39 (0.12)	-0.34 (0.22)
3	-1.84 (0.06)	-1.12 (0.03)	-0.76 (0.02)	-0.47 (0.03)	-0.33 (0.07)
4	-1.34 (0.03)	-0.73 (0.02)	-0.45 (0.04)	-0.37 (0.06)	-0.11 (0.09)
5	-1.94 (0.08)	-1.26 (0.03)	-0.84 (0.02)	-0.49 (0.03)	-0.19 (0.09)
6	-1.47 (0.03)	-0.80 (0.02)	-0.44 (0.03)	-0.43 (0.04)	-0.09 (0.07)
7	-1.47 (0.05)	-1.06 (0.03)	-0.81 (0.03)	-0.52 (0.03)	-0.43 (0.06)
8*	-1.20 (0.02)	-0.49 (0.03)	-0.10 (0.05)	-0.09 (0.07)	-0.25 (0.11)
9*	-1.55 (0.05)	-0.93 (0.03)	-0.67 (0.03)	-0.40 (0.05)	-0.42 (0.11)
10	-1.73 (0.05)	-1.06 (0.03)	-0.60 (0.02)	-0.54 (0.04)	-0.44 (0.09)
11	-1.31 (0.03)	-0.69 (0.02)	-0.21 (0.03)	-0.10 (0.09)	-0.08 (0.09)
12	-1.61 (0.05)	-1.05 (0.03)	-0.71 (0.03)	-0.56 (0.03)	-0.38 (0.05)

* Las medias de θ no aumentan ante valores crecientes de la escala Likert.

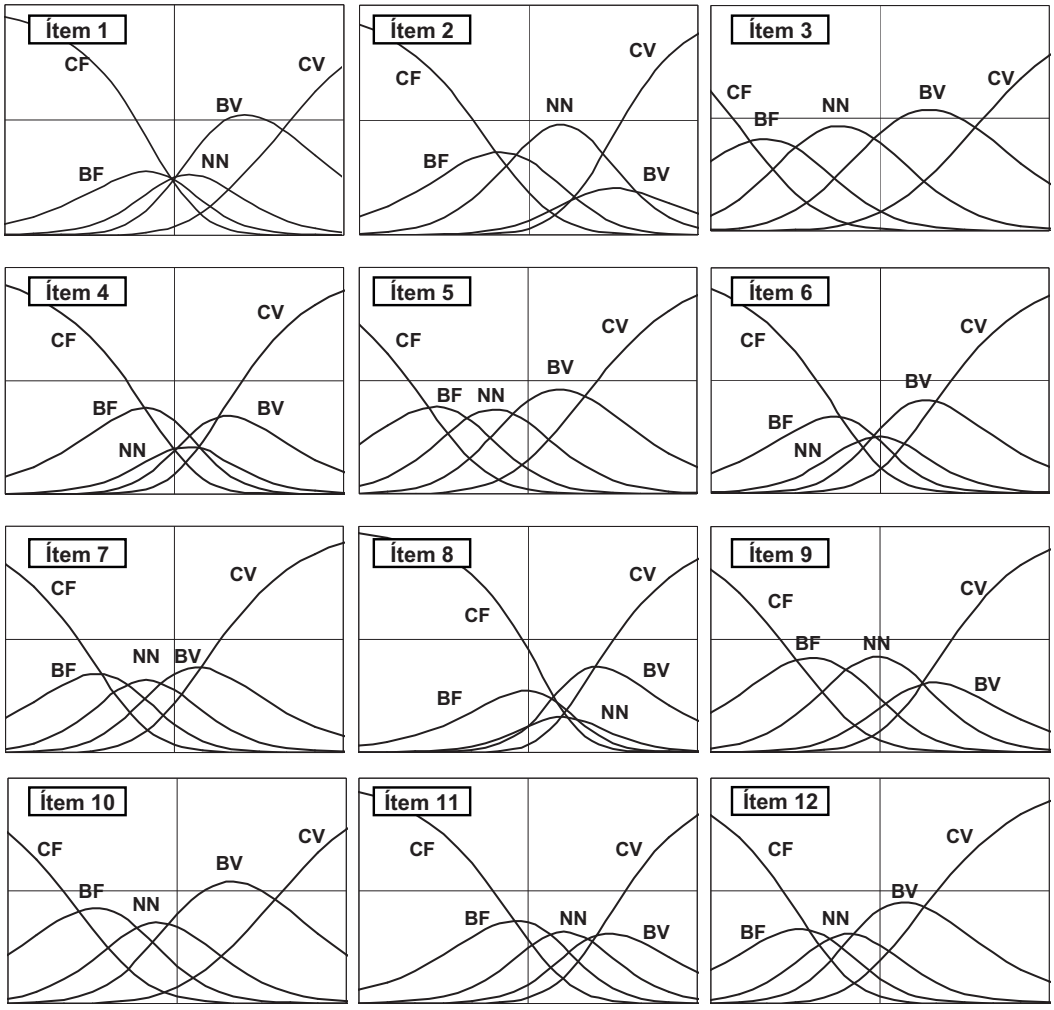

Figura 2. Curvas Características de las Categorías de Respuesta de los ítems. CV = Completamente verdadero para mí; BV = Bastante verdadero para mí; NN = Ni verdadero ni falso para mí; BF = Bastante falso para mí; CF = Completamente falso para mí.

en el orden de los parámetros β_{ih}. El caso más extremo de este comportamiento inadecuado aparece en el ítem 8, en el que hay dos categorías (*Bastante falso para mí y Ni verdadero ni falso para mí*) que no resultan máximamente probables en ninguna región del continuo del rasgo.

Comentarios

Este trabajo busca contribuir con el avance de un área en incipiente desarrollo en la región como es la aplicación de modelos politómicos de la TRI a tests de comportamiento típico. Siguiendo esta idea rectora, se aportó una formulación detallada del MCP para mejorar su compresión matemática y se efectuó un análisis exhaustivo que de las propiedades psicométricas de los ítems de la escala Distorsión del BFQ.

Como se ha podido apreciar, los 12 ítems que componen el instrumento han mostrado un ajuste aceptable al MCP. Sin embargo, para datos provenientes de tests de comportamiento típico, este ajuste es un requisito necesario pero no suficiente si se pretende dar garantías del adecuado funcionamiento del ítem. El MCP no condiciona la estimación de los parámetros de localización con valores ordenados. De esta manera, el modelo contempla la aparición de β_h invertidos aunque, como se mencionó anteriormente, también respeta que los pasos están siempre secuenciados. En un ítem de ejecución típica cuya resolución requiere de varios pasos sucesivos podría resultar que el segundo paso demande un nivel menor del rasgo que el primer paso. Por ende, el β_h del segundo paso resultaría menor al β_h del primer paso. No obstante, el modelo supone que se enfrentan al segundo paso solo quienes han superado el primer paso (Embretson & Reise, 2000). Esta interpretación no es aplicable a formatos de respuesta Likert porque los anclajes lingüísticos de las categorías y los numerales usados para su codificación se determinan considerando la propiedad de orden creciente. Por ende, la inversión de los parámetros muestra que las diferentes categorías de la escala Likert no reflejan crecientes niveles del rasgo latente.

La inversión del orden de los β_h en cinco reactivos de la escala Distorsión señala que el sistema utilizado para recoger las respuestas a estos ítems no cumple con los requisitos necesarios para una medición válida. Lo recomendado habitualmente, en el análisis de ítems con estas características, es revisar los reactivos y el formato de respuesta desde una perspectiva sintáctica y semántica (Andrich, 2010; Rojas & Pérez, 2001). En primera instancia, deberían ser eliminados o bien, en el caso de no hallar una justificación para el comportamiento inadecuado, deberían ser sustituidos. Se observó que la mayoría de los ítems cuyos parámetros presentaron inversiones parecen reflejar problemas en la manera en que los evaluados interpretan la categoría central de la escala Likert. Esto conduce a cuestionar la implementación de una escala con cinco opciones o a reconsiderar la eficacia del anclaje lingüístico Ni verdadero ni falso para mí. Sin embargo, este análisis no es extensivo al resto de los ítems en donde la categoría central se muestra como máximamente probable en algún rango del rasgo latente.

Ante este hallazgo, no parece conveniente analizar la precisión de la medida mediante la Función de Información del Test (FIT) (Martínez Arias, 1995). Como demostraron Dodd y Koch (1987), la FIT se eleva en niveles del rasgo en donde aparecen parámetros β_h con orden inverso. En consecuencia, cabe esperar que el nivel de información disminuya frente a la posibilidad de reformulación o eliminación de estos ítems.

En cuanto al estudio del ajuste de los sujetos al MCP, se ha observado que una proporción de evaluados no presentaron un patrón de respuestas acorde con el propuesto por el modelo. Aunque considerable, este porcentaje de personas que no ajustan a las predicciones del modelo es similar a la reportadas por estudios análogos con tests de comportamiento típico (Curtis, 2004; Rojas & Pérez, 2001). Las posibles causas que originan este desajuste son variadas y deberían identificarse al estudiar detenidamente caso por caso. Se trata de fuentes de variabilidad ajenas al constructo evaluado pero que han tenido incidencia en la medición de algunos sujetos. Tradicionalmente, se menciona a la deseabilidad social como uno de los factores más importantes que explican el desajuste pero, por obvias razones, esta justificación no incumbe

a los resultados del presente estudio. Según Curtis (2004) y Karabatsos (2000), otras condiciones que podrían explicar patrones desajustados de respuesta son la falta de cooperación o motivación, problemas de distracción o compresión lectora del evaluado. Sin embargo, estos aspectos fueron controlados durante el procedimiento de recolección de los datos.

Un punto para destacar, en tanto que podría considerarse como una limitación del presente estudio, es que la recolección de los datos se realizó en condiciones de administración neutrales o de baja motivación para la distorsión. La escala Distorsión asume que la tendencia a ofrecer un perfil falseado de manera más o menos intencionada es un rasgo relativamente estable. Pero es sabido que la utilidad práctica de la escala corresponde a su aplicación en contextos organizacionales o jurídicos donde las decisiones surgidas de la evaluación tienen importantes consecuencias para los evaluados.

Las condiciones en que se realizaron las administraciones en este estudio podrían tener implicaciones en la métrica del rasgo latente. No existe una manera particularmente recomendable para establecer esta métrica. Los estudios instrumentales que modelizaron con TRI los ítems de deseabilidad social partieron de datos obtenidos en condiciones neutras de motivación para la disimulación (Ferrando & Chico, 2000; Seol, 2007; Vésteinsdóttir et al., 2017). Sin embargo, las aplicaciones en determinados contextos podrían modificar considerablemente los valores estimados para los parámetros de los ítems. Es por esta razón que podría resultar de interés en estudios futuros analizar el comportamiento de la prueba en situaciones donde varíen la motivación del sujeto para responder y los intentos por manipular la información aportada.

Al hilo de lo anterior, el rango de adhesión de los ítems entre -2.30 y 1.90 resultó más acotado que el esperable (i.e. -3 y 3), considerando que son las puntuaciones extremas de la escala las que se interpretan para determinar el grado de distorsión aceptable. Aun cuando se está modelizando el instrumento a partir de una muestra de personas con baja motivación para el falseamiento, se esperaría que algunos niveles de adhesión de los ítems también tiendan a localizarse en los extremos del rasgo.

Otro resultado para resaltar es la concentración de la distribución de los θ de los participantes en valores medio-bajos del rasgo, cuando no estaban dadas las circunstancias para que los participantes tiendan a ofrecer un perfil distorsionado en un sentido negativo. Esto podría indicar que la escala presenta dificultades en la determinación de la bipolaridad que se le atribuye al constructo.

Valores bajos en la escala podrían reflejar simplemente una escasa predisposición a ofrecer una imagen positiva de sí mismo en lugar de pretender mostrar una imagen negativa. Los contenidos de los ítems de las pruebas que estudian la deseabilidad social se redactan a partir de identificar comportamientos deseables pero improbables (estar de acuerdo con los demás en todo momento) o indeseables pero presentes con frecuencia (haber desobedecido órdenes en la infancia). Pero el grado en que cada una de estas conductas son deseables y frecuentes se define en función de la muestra normativa (Burns & Christiansen, 2006), por lo que, presumiblemente, habría que refinar los indicadores usados en la construcción de los ítems.

En conclusión, los resultados obtenidos en la modelización con el MCP revelaron que, en su forma actual, la escala de Distorsión adolece de problemas que cuestionan la validez de los puntajes. Resulta indispensable verificar una adecuada relación entre el continuo del rasgo latente y la escala ordinal empleada como formato de respuesta de los ítems. Estas mejoras podrían redundar en una reducción del porcentaje de personas con patrones de respuesta desajustados. Las sugerencias realizadas demandan de una nueva administración del instrumento a fin de ensayar y avalar empíricamente las modificaciones introducidas. Futuras investigaciones se encaminarán sobre estos objetivos.

Referencias

Abad, F., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Abal, F., Galibert, M., Aguerri, M., & Attorresi, H. (2014). Comparación de los Modelos Respuesta Graduada y Crédito Parcial aplicados a una escala de Utilidad de la Matemática. *Revista Argentina de Ciencias del Comportamiento, 6*(3), 6-16.

Abal, F., Lozzia, G. Aguerri, M., Galibert, M., & Attorresi, H. (2010). La escasa aplicación de la Teoría de Respuesta al Ítem en tests de ejecución típica. *Revista Colombiana de Psicología, 19*(1) 111-122.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. doi: 10.1177/0013164411432166

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 123-152). Mahwah, NJ: Erlbaum.

Andrich, D. (2013). An Expanded Derivation of the Threshold Structure of the Polytomous Rasch Model That Dispels Any "Threshold Disorder Controversy". *Educational and Psychological Measurement, 73*(1), 78–124. doi: 10.1177/0013164412450877

Asgeirsdottir, R. L., Vésteinsdóttir, V. &, Thorsdottir, F. (2016). Short form development of the Balanced Inventory of Desirable Responding: Applying confirmatory factor analysis, item response theory, and cognitive interviews to scale reduction *Personality and Individual Differences, 96*, 212–221. doi: 10.1016/j.paid.2016.02.083

Ato, M., López, J. J., & Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anales de Psicología, 29*(3), 1038 - 1059. doi: 10.6018/analesps.29.3.178511

Barbaranelli, C., & Caprara, G. V. (2002). Studies of the big five questionnaire. In B. de Raad & M. Perugini (Eds). *Big five assessment* (pp. 109-124). Ashland, OH, US: Hogrefe & Huber Publishers.

Bermúdez, J. (1995). *Manual del Cuestionario "Big Five" (BFQ)*. Madrid: TEA.

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2nd Ed.). New York: Guilford Press.

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R. L. Griffith, & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113-148). Greenwich, CT: Information Age Publishing.

Byrne, B. M. (2012). S*tructural equation modeling with Mplus: Basics, concepts, applications, and programming*. New York: Routledge.

Caprara, G. V., Barbaranelli, C., Bermúdez, J., Maslach, C., & Ruch, W. (2000). Multivariate methods for the comparison of factor structures in cross-cultural research. An illustration with the Big Five Questionnaire. *Journal of Cross-cultural Psychology, 31*(4) 437-464. doi: 10.1177/0022022100031004002

Caprara, G. V., Barbaranelli, C., & Borgogni, L. (1993). *Big Five Questionnaire (BFQ). Manuale*. Florencia: Organizzazioni Speciali.

Caprara, G. V., Barbaranelli, C., & Borgogni, L. (1996). Big Five Questionnaire (BFQ): Caratteristiche psicometriche e validità transculturale. *Archivio di Psicologia, Neurologia e Psichiatria, 57*, 486-504.

Carretero-Dios, H., & Pérez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *International Journal of Clinical and Health Psychology, 7*(3), 863-882.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354. doi: 10.1037/h0047358

Curtis, D. D. (2004). Person Misfit in Attitude Surveys: Influences, Impacts and Implications. *International Education Journal, 5*(2), 125-144.

DiStefano, C., Morgan, G. B., & Motl, R. W. (2012). An examination of personality characteristics related to acquiescence. *Journal of applied measurement, 13*(1), 41-56.

Dodd, B. G., & Koch, W. R. (1987). Effects of Variations in Item Step Values on Item and Test Information in the Partial Credit Model. *Applied Psychological Measurement, 11*(4), 371-384. doi: 10.1177/014662168701100403

Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum Publishers.

Ferrando, P., & Chico, E. (2000). Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne. *Psicothema, 12*(3), 383-389.

García-Pérez, M. A. (2017). An Analysis of (Dis) Ordered Categories, Thresholds, and Crossings in Difference and Divide-by-Total IRT Models for Ordered Responses. *The Spanish Journal of Psychology, 20*(10), 1–27. doi: 10.1017/sjp.2017.11

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152-176.

Kline, R. B. (2011). *Principles and practice of structural equation modelling* (3rd. Ed.). New York: Guilford Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2006). *Winsteps*® (Version 3.63.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monograph Nº7. Iowa City, IA: Psychometric Society.

Martínez Arias, M. R. (1995). *Psicometría: Teoría de los Tests Psicológicos y Educativos*. Madrid: Síntesis.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149 - 174. doi: 10.1007/BF02296272

Masters, G. N. (2016). Partial Credit Model. En W. J. van der Linden (Ed.). *Handbook of Item Response Theory, Volume 1: Models* (pp. 109-126). Boca Raton: Chapman & Hall/CRC.

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. En W. J. Van der Linden y R. K. Hambleton (Eds.). *Handbook of Modern Item Response Theory*, (pp. 101-121). New York: Springer.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward Modern Psychometrics. Application of Item Response Theory Models in Personality Research. In R. W. Robins, R. C. Fraley & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology*, (pp. 407 - 423). New York: Guilford Press.

Muthén, L., & Muthén, B. (2010). *Mplus User's Guide, 6th Edn*. Los Angeles, CA: Muthén & Muthén.

Panter, A. T., Swygert, K. A., & Dahlstrom, G. W. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68*, 561-589. doi: 10.1207/s15327752jpa6803_6

Paulhus, D. L. (1988). *Manual for the Balanced Inventory of Desirable Responding* (BIDR-6). Manuscrito inédito, Universidad de Columbia Británica.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhague: The Danish Institute for Educational Research.

Reise, S. P., & Revicki, D. (2015). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48. doi: 10.1146/annurev.clinpsy.032408.153553

Rojas, A. J., & Pérez, C. (2001) *Nuevos Modelos para la Medición de Actitudes*. Valencia: Promolibro.

Seol, H. (2007). A psychometric investigation of the Marlowe-Crowne social desirability scale using Rasch measurement. *Measurement and Evaluation in Counseling and Development, 40*, 155-168.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: a review. *Assessment, 18*(3), 291-307. doi: 10.1177/1073191110374797

Vendramini, C., Silva, M., & Dias, A. (2009). Avaliação de atitudes de estudantes de psicologia via modelo de crédito parcial da TRI. *Psico-USF, 14*(3), 287-298. doi: 10.1590/S1413-82712009000300005

Vésteinsdóttir, V., Reips, U., Joinson, A., & Thorsdottir, F. (2017). An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews. *Personality and Individual Differences, 107*, 164–173. doi: 10.1016/j.paid.2016.11.023

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the Balanced Inventory of Desirable Responding: Dichotomous versus polytomous conventional and IRT scoring. *Psychological Assessment, 26*(3), 878-891. doi: http://dx.doi.org/10.1037/a0036430

Wetzel, E., & Carstensen, C. H. (2014). Reversed Thresholds in Partial Credit Models. A Reason for Collapsing Categories? *Assessment, 21*(6), 765-774. doi: 10.1177/1073191114530775

Wrigth, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: Mesa Press.

Recibido: 9 de marzo de 2016
Aceptado: 3 de mayo de 2017