Psicodebate, Vol. 19, Nº 1, Junio 2019 - Diciembre 2019
ISSN: 1515-2251
e-ISSN: 2451-6600
31 – 50
http://dx.doi.org/10.18682/pd.v1i1.854

Construcción de un banco de ítems de facetas de neuroticismo para el desarrollo de un test adaptativo

Facundo Juan Pablo Abal[1], Sofía Esmeralda Auné[1] y Horacio Félix Attorresi[2]

RESUMEN

El objetivo de este trabajo fue elaborar un banco de ítems para medir las facetas del Neuroticismo basado en el Modelo de los Cinco Factores (McCrae & Costa, 2010) y examinar la viabilidad de una administración adaptativa. Se inició con un pool de 90 ítems, que fue reducido a 54 (nueve por faceta) por juicio experto y estudio piloto. La versión depurada se administró a 1147 adultos de población general del área metropolitana de Buenos Aires (52.7% mujeres). Un 70% de los casos se usó para: a) calibrar los ítems de cada faceta por separado con el Modelo de Respuesta Graduada de Samejima (2010), b) estudiar la estructura interna del banco con un Análisis Factorial Confirmatorio y c) obtener evidencias de validez concurrente. El alfa ordinal de las facetas osciló entre .76 y .87. Con el 30% restante de casos se efectuó una simulación de administración adaptativa con criterio de parada de error ≤ 0.50. Se concluye que el banco reúne evidencias de validez y confiabilidad aceptables para su administración en un formato convencional, pero se necesita incorporar más ítems si se pretende optimizar la medición de las facetas Impulsividad, Vulnerabilidad y Hostilidad.

Palabras clave: neuroticismo, modelo de lo cinco factores, banco de ítems, test adaptativo informatizado, teoría de respuesta al ítem.

Constructing a bank of neuroticism facet items for the development of an adaptive test

ABSTRACT

The goal of this work was to elaborate a bank of items to measure the facets of neuroticism on the basis of the Five-Factor Model (McCrae & Costa, 2010) and the feasibility of adaptive administration was examined. The study began with a pool of 90 items which was then reduced to 54 (nine per facet) through expert judgment and a pilot study. The refined version was administered to 1147 adults in the general population of Buenos Aires metropolitan area (52.7% women). Seventy percent of the cases were used to: a) calibrate the items of each facet separately with Samejima's (2010) Graded Response Model, b) verify internal structure of bank through Confirmatory Factorial Analysis and c) obtain evidence of concurrent validity. Facets ordinal alpha ranged between .76 and .87. With the remaining 30% of cases, an adaptive administration simulation was carried out considering an estimation error ≤ 0.50 as stopping criteria. It is concluded that the bank collects evidence of acceptable validity and reliability for its administration in a conventional format but requires the incorporation of more items to optimize the Impulsiveness, Vulnerability and Angry Hostility facets measurement.

Keywords: neuroticism, five factor model, item bank, computerized adaptive test, item response theory.

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires, Argentina; fabal@psi.uba.ar
[2] Universidad de Buenos Aires, Argentina.

Existe un consenso extendido en el marco de las teorías de los rasgos para entender al Neuroticismo como una de las dimensiones fundamentales en la estructura de la personalidad normal. Su importancia ya había sido señalada tempranamente en teorías factorialistas clásicas como las de Cattell y Guilford y en el modelo psicobiológico de Eysenck. Pero su relevancia se ha consolidado con la hegemonía alcanzada en las últimas décadas por el Modelo de los Cinco Factores -*Five Factor Model*, FFM- (Digman, 1990; Goldberg, 1993, 1999; McCrae & Costa, 2010). Es justamente en el marco de esta teoría en donde el Neuroticismo también ha sido denominado con frecuencia a partir de su polo opuesto, la Estabilidad Emocional (Caprara, Barbaranelli, Borgogni, & Vecchione, 2007; Goldberg, 1999; Norman, 1963).

El Neuroticismo describe la tendencia de una persona a experimentar afectos negativos de manera intensa y/o recurrente frente a diferentes fuentes de estrés. Entre los afectos más recurrentemente mencionados en la literatura suelen aparecer los sentimientos de miedo, tristeza, culpa y enojo. La inclusión de otros aspectos, tales como los sentimientos de inferioridad y dependencia o comportamientos impulsivos, son discutidos por los teóricos de la personalidad normal (Jeronimus, Kotov, Riese, & Ormel, 2016; Widiger, 2009). Los sujetos con niveles elevados de Neuroticismo perciben el mundo como un lugar amenazante y se creen con dificultades para afrontar eventos desafiantes. El otro extremo de esta dimensión, un nivel bajo de Neuroticismo, describe a las personas que tienden a permanecer serenas y relajadas incluso ante situaciones que podrían provocar tensión (Goldberg, 1993).

La complejidad conceptual del Neuroticismo ha llevado a los autores del FFM a identificar un conjunto de facetas anidadas jerárquicamente en este dominio. Estas facetas discriminan la variedad de emociones y sentimientos negativos concebidos como parte del Neuroticismo y permiten reflejar cierto grado de heterogeneidad en la descripción de las diferencias individuales del dominio. Aunque resultan insoslayables los desacuerdos sobre la cantidad y especificidad de las facetas (Tackett & Lahey, 2017; Watson, Nus, & Wu, 2017) la propuesta empírico-racional de McCrae y Costa (2003, 2010) es la más replicada en estudios transculturales y cuenta con el apoyo de otros investigadores del FFM (Goldberg, 1999; Johnson, 2014). Según McCrae y Costa (2003), cada una de las seis facetas definidas para el Neuroticismo está determinada por algún tipo de emoción o sentimiento negativo que le brinda entidad. Las facetas Ansiedad y Hostilidad ponen de manifiesto la tendencia a experimentar estados emocionales de temor y enojo. Los sentimientos de tristeza y vergüenza son los pilares sobre los que se asientan las facetas Depresión y Autoconciencia. Finalmente, las facetas Impulsividad y Vulnerabilidad responden más a un orden comportamental. La primera se distingue por la imposibilidad de controlarse en momentos de arrebato y la segunda por la tendencia a inhibirse ante situaciones de estrés.

Aun cuando el Neuroticismo es un rasgo de la personalidad normal, la evidencia empírica ha demostrado que causa un fuerte impacto en los sistemas de salud (Hajek, Bock, & König, 2017; Vittengl, 2017; Widiger & Oltmanns, 2017). Se trata de un factor de vulnerabilidad subyacente para el desarrollo y mantenimiento de distintos trastornos psicopatológicos y enfermedades físicas como cardiopatías, diabetes, asma o síndrome de intestino irritable (Lahey, 2009). Pero el Neuroticismo también ofrece como ventaja la posibilidad de pensar estrategias de intervención y prevención transdiagnósticas. Investigaciones actuales sostienen que este rasgo puede

resultar más maleable de lo que se suponía anteriormente, lo que ha motivado el diseño de tratamientos con objetivos centrados en la disminución del Neuroticismo (e.g. Drake, Morris, & Davis, 2017; Sauer-Zavala, Wilner, & Barlow, 2017). En línea con esta perspectiva, se ha comenzado a recomendar la detección de niveles altos de Neuroticismo en población general durante la asistencia clínica de rutina (e.g. Hengartner, Kawohl, Haker, Rössler, & Ajdacic-Gross, 2016; Widiger, 2009; Widiger & Oltmanns, 2017).

Ante estos objetivos evaluativos, propios del ámbito clínico-epidemiológico, son apreciados especialmente los instrumentos cortos que puedan ofrecer resultados válidos y confiables para una gran cantidad de sujetos en el menor tiempo posible (Baldasaro, Shanahan & Bauer, 2013). Sin embargo, los inventarios de personalidad más destacados como NEO-PI-3 (McCrae & Costa, 2010) o NEO-IPIP (Goldberg et al., 2006) destinan 48 y 60 ítems respectivamente para evaluar las seis facetas de Neuroticismo. Son tests que brindan una información pormenorizada de las facetas, pero que resultan poco prácticos por su extensión. Una alternativa son las escalas breves (e.g. NEO-FFI; McCrae & Costa, 2010) que reducen la cantidad de ítems porque permiten una valoración unidimensional y excesivamente global del rasgo. Otros tests que operacionalizan el FFM logran disminuir el conjunto de elementos para Neuroticismo eliminando algunas facetas, fusionándolas en una menor cantidad de subdimensiones o haciendo más homogéneo el contenido de los ítems (Caprara et al., 2007; Soto & John, 2017b; Taylor & DeBruin, 2006). También se han desarrollado pruebas cortas y extra-cortas (Donnellan, Oswald, Baird, & Lucas, 2006; Gosling, Rentfrow & Swann, 2003; Soto & John, 2017a) con el objeto de disminuir los errores de medida provocados por fatiga o aburrimiento del evaluado.

Ahora bien, los investigadores deberían tener en cuenta que la ganancia práctica que brindan los cuestionarios más cortos se obtiene a expensas de resignar una cuantía de calidad psicométrica (Credé, Harms, Niehorster, & Gaye-Valentine, 2012). Cuando el constructo tiene una estructura interna compleja, como ocurre con el Neuroticismo, los ítems que componen una prueba breve suelen presentar correlaciones modestas que reflejan la relativa heterogeneidad del contenido. Como resultado, disminuyen los coeficientes de consistencia interna usados para estimar la confiabilidad (Baldasaro et al., 2013; Sibley, 2012). Por otra parte, una estrategia aplicada con frecuencia para reducir la cantidad de ítems es conservar solo los elementos con elevada capacidad discriminativa para aumentar la consistencia interna. Esto tiene como consecuencia un riesgo potencial sobre la representatividad y exhaustividad del contenido (Milojev, Osborne, Greaves, Barlow, & Sibley, 2013; Morizot, 2014; Ziegler, Kemper & Kruyen, 2014). Las variaciones en la cobertura podrían incluso originar diferencias entre las mediciones obtenidas con los distintos instrumentos que cuantifican el Neuroticismo (Ormel et al., 2013). En suma, las estrategias usadas para construir escalas breves se enfrentan a la búsqueda de un equilibrio entre la precisión de la medida y la validez de contenido.

En los últimos años ha crecido la cantidad de estudios instrumentales sobre Test Adaptativos Informatizados (TAIs) desarrollados en el marco de la Teoría de Respuesta al Ítem (TRI) como una estrategia metodológica para acortar los tests de personalidad sin comprometer la confiabilidad o validez de la medida (Reise & Revicki, 2015). Un TAI permite realizar una evaluación más eficiente seleccionando progresivamente los ítems que aportan más

información en función de las respuestas que va manifestando el individuo. La administración continúa hasta que se consigue una cantidad especificada de ítems o se alcanza un valor prefijado de error estándar. Todo TAI se basa en un banco de ítems relativamente amplio que ha sido calibrado antes con un modelo de la TRI. Esto implica que son conocidas las propiedades psicométricas de los ítems y que se cumplen los supuestos de dimensionalidad e independencia local que requiere el ajuste del modelo utilizado (Olea & Ponsoda, 2013).

Se han llevado adelante algunos estudios de versiones adaptativas usando como banco los cuestionarios de personalidad tradicionales validados en el marco de la Teoría Clásica de Tests (e.g. Forbey & Ben-Porath, 2007; Reise & Henson, 2000). Pero los ítems empleados para los TAIs demandan análisis previos cualitativos y cuantitativos más rigurosos que no siempre están garantizados en la validación desde la perspectiva clásica (Abal, Lozzia, Aguerri, Galibert & Attorresi, 2010; DeWalt, Rothrock, Yount & Stone, 2007; Reise & Revicki, 2015). Siguiendo esta línea, algunos de los TAIs desarrollados más recientemente elaboran ítems originales como punto de partida. Nieto, et al. (2017) ensayaron la evaluación adaptativa de 360 ítems nuevos de un banco que mide la personalidad según el FFM y encontraron que se podían obtener estimaciones aceptables usando solo cuatro ítems por faceta. Drasgow et al. (2012) optaron por diseñar el *Tailored Adaptive Personality Assessment System* (TAPAS) con ítems de elección forzada para evitar distorsiones intencionales en las respuestas. Estos autores reportaron una disminución del 50% de los ítems administrados al simular una aplicación adaptativa. Makransky, Mortensen y Glas (2013), en cambio, utilizaron los ítems originales del NEO-PI-R, pero aplicando un TAI multidimensional. Ellos demostraron que, con este procedimiento, la estimación de los puntajes se puede realizar de manera más eficiente en las facetas que correlacionaban más fuertemente.

La medición del Neuroticismo y sus facetas se encuentra fuertemente arraigada en formas de administración convencionales con cuestionarios compuestos por una cantidad fija de ítems, que suelen responderse en formato de lápiz y papel. A causa de la vinculación con variables salugénicas y psicopatológicas, puede resultar de utilidad contar con un instrumento que permita la medición eficiente de las faceras de este dominio y que, al mismo tiempo, se adecúe a las condicionantes que impone la evaluación en un escenario clínico-epidemiológico. El ahorro de tiempo que supondría la implementación de un TAI podría ser fructífero para la evaluación de otras variables que resulten de interés o, simplemente, para brindar condiciones de mayor bienestar al evaluado durante la administración.

Por consiguiente, en este trabajo se propone como objetivo general presentar la primera etapa de la construcción de un banco de ítems para la medición de las facetas del Neuroticismo. Los objetivos específicos consisten en: a) calibrar seis conjuntos de ítems que operacionalicen las facetas del Neuroticismo según la taxonomía de McCrae y Costa (2010), y b) examinar la viabilidad de una aplicación adaptativa de estos ítems tendiente a minimizar la cantidad de elementos necesarios para su medición.

MÉTODO

Participantes

Participaron 1147 adultos de población general residentes en el área metropolitana de Buenos Aires, Argentina. Los sujetos se seleccionaron a partir de un muestreo no probabilístico por accesibilidad. La edad media era de 29.7 años (*DE* = 11.9; *Mín* = 18, *Máx* = 82) y el 52.7% consignó pertenecer al género femenino. Solo un 5.2% no alcanzó a concluir el nivel educativo medio. El resto de los participantes completaron estudios secundarios (56.2%), terciarios (15.5%) y universitarios (23.1%). En función del lugar de residencia y el nivel de estudio alcanzado por los participantes se podría considerar de manera estimativa que la mayoría tienen un nivel socioeconómico medio con una menor proporción de medio-bajo y medio-alto.

Instrumentos

Banco de ítems para las Facetas de Neuroticismo. Se llevó adelante una revisión de elementos procedentes de diversos instrumentos reconocidos que miden la dimensión Neuroticismo, sus facetas y constructos emparentados conceptualmente. Los contenidos de los indicadores empíricos recolectados fueron usados como fuentes para la redacción de ítems originales que se ajustan a la composición de facetas propuestas por McCrae y Costa (2010).

Con el objetivo de hacer operativa la construcción del banco, en esta etapa se partió de una selección de 90 ítems (15 por faceta) y se reservó el resto de los ítems para futuras calibraciones. La selección de estos 90 ítems se realizó mediante un análisis cualitativo del contenido: a) se apartaron enunciados que resultaban similares a ítems pertenecientes a los cuestionarios EPQ-RS y SCL-90-R, dado que en el presente estudio ambos instrumentos fueron usados exclusivamente para obtener evidencias de validez concurrente, y b) se corroboró que los contenidos no fueran redundantes entre sí para evitar la violación del supuesto de independencia local de los ítems durante el análisis con TRI (DeWalt et al., 2007; Reise & Revicki, 2015).

Una depuración posterior realizada por la crítica de jueces expertos redujo los 90 ítems iniciales a 54 (nueve por faceta). De esta manera quedó conformada la versión administrada del banco. Si bien la mayoría de los ítems del banco son originales entre los 54 ítems analizados, quedaron incluidos 18 enunciados pertenecientes (tres por faceta) a la adaptación argentina del Inventario IPIP-NEO (Cupani, Pilatti, Urrizaga, Chincolla, & Richaud de Minzi, 2014) del Banco internacional de ítems de Personalidad (IPIP) de Goldberg et al. (2006). La incorporación de estos ítems tuvo como finalidad anticipar la aparición de la estructura factorial del constructo desde el diseño del banco, utilizando elementos característicos de cada faceta que han demostrado un adecuado funcionamiento a nivel local.

Con respecto al formato de respuesta, se optó por una escala tipo Likert con cuatro categorías (*En desacuerdo*, *Ligeramente en desacuerdo*, *Ligeramente de acuerdo* y *De acuerdo*). Este formato también se usó para los 18 ítems del IPIP-NEO para uniformizar la estructura del cuestionario. La decisión referida a este diseño se basa en lineamientos derivados de estudios que mostraron que cuatro opciones puede resultar una cantidad óptima para garantizar un equilibrio entre el grado de ajuste del modelo de la TRI y la

precisión de la medida (e.g., Abal, Auné, Lozzia & Attorresi, 2017; Lozano, García-Cueto & Muñiz, 2008).

Eysenck Personality Questionary revised short version, EPQ-RS (Eysenck & Eysenck, 1994; adaptación de Squillace, Picón-Janeiro & Schmidt, 2013). Se compone de 42 ítems redactados en forma de pregunta y con formato de respuesta dicotómico (Sí-No). Los adaptadores del instrumento replicaron la estructura de tres factores propuestos por el modelo de Eysenck (Psicoticismo, Extraversión y Neuroticismo) y un cuarto factor usado para evaluar la sinceridad en las respuestas. Los estudios de confiabilidad de las cuatro escalas registraron índices de consistencia interna adecuados (KR-20 entre .66 y .84). En la muestra total del presente estudio el análisis de consistencia interna mostró valores de KR-20 similares aunque ligeramente más elevados (entre .69 y .86).

Inventario de síntomas SCL-90-R (Derogatis, 1994). Consta de 90 ítems que indagan sobre la intensidad con la que se han experimentado diversos síntomas psicológicos durante los últimos siete días. Tiene un formato de respuesta politómica con cinco opciones (*Nada*, *Muy poco*, *Poco*, *Bastante* y *Mucho*). Los ítems se agrupan para posibilitar la medición de nueve dimensiones clínicas (Somatización, Obsesiones y Compulsiones, Sensitividad Interpersonal, Depresión, Ansiedad, Hostilidad, Ansiedad Fóbica, Ideación Paranoide y Psicoticismo) y tres índices globales (Índice de Severidad Global, Total de Síntomas Positivos e Índice de Malestar Positivo). La adaptación local muestra evidencias de validez y estudios de confiabilidad adecuados tanto para población clínica (Sánchez & Ledesma, 2009) como no clínica (Casullo, 2004). La consistencia interna de todos los ítems del inventario mostró un Alfa de Cronbach de .96 en la muestra total de este estudio, mientras que para las dimensiones clínicas este coeficiente osciló entre .77 y .86.

Procedimiento y análisis de datos

Depuración: Juicio experto y prueba piloto. Los 90 ítems propuestos inicialmente para el banco de facetas del Neuroticismo fueron revisados a ciegas por cinco jueces. Se solicitó a este panel de expertos que valoren la congruencia del enunciado con la definición de la faceta y la relevancia del indicador. Cada uno de estos aspectos debía ser evaluado usando una escala graduada de tres categorías. Se utilizó el índice V de Aiken (Penfield & Giacobbi, 2004) para analizar el acuerdo entre los expertos. Durante esta fase se conservaron los 9 ítems de cada faceta que mostraron mayor consenso entre los jueces. Para la valoración de la congruencia con la definición conceptual se adoptó como criterio de aceptación un límite inferior del intervalo de confianza del V de Aiken ≥ .60 para un nivel de confianza del 90%. En cambio, para analizar la relevancia del indicador se tomó un criterio más liberal (límite inferior IC del V de Aiken ≥ .50), debido a la dificultad que supone encontrar un muestreo de conductas que no resulte redundante en facetas cuyas definiciones conceptuales son más acotadas (Reise & Rodríguez, 2016). La información proveniente del juicio experto fue complementada con un estudio piloto (n = 35) a fin de revisar aspectos formales y generar evidencias de validez aparente.

Recolección de datos. Los participantes contestaron el protocolo en formato lápiz-papel de manera individual y sin tiempo límite. Las administraciones fueron realizadas por psicólogos y alumnos avanzados de la carrera de Psicología debidamente entrenados y supervisados. Previa

aplicación, se explicó a los evaluados que la investigación tenía como finalidad la medición de atributos de su personalidad y que no había respuestas correctas o incorrectas. Se los informó sobre el carácter voluntario de su participación y la posibilidad de abandonar la evaluación en cualquier momento de la actividad. También se les comunicó sobre las garantías de anonimato y confidencialidad de sus respuestas. Todas estas condiciones fueron explicitadas además en la redacción del consentimiento que debieron firmar los sujetos para participar.

Calibración de los ítems con TRI. Dado que usar la misma muestra para calibrar los ítems y para simular el TAI puede llevar a resultados espurios (Smits, Cuijpers, & van Straten, 2011), se dividió a los participantes en dos grupos de manera aleatoria. Las respuestas de un 70% de los sujetos (n = 798) fueron empleadas para calibrar los ítems según el modelo de respuesta graduada de Samejima (2010), mientras que el resto de los individuos (n = 349) fueron considerados exclusivamente para analizar el TAI.

Se realizó un análisis factorial confirmatorio (AFC) para cada faceta con el objetivo de verificar la unidimensionalidad requerida por el modelo de la TRI. Se utilizó el programa Mplus (Muthén & Muthén, 2010) para estimar los parámetros con el método robusto de mínimos cuadrados ponderados (*Weighted Least Squares Mean and Variance Adjusted*, WLSMV) sobre la base de la matriz de correlaciones policóricas. El ajuste del modelo a los datos se analizó con los criterios recomendados por Byrne (2012): índices de ajuste comparativo *CFI* ≥ .90 y de Tucker-Lewis *TLI* ≥ .90 y el error medio cuadrático de aproximación RMSEA ≤ .08.

Se aplicó el Modelo de Respuesta Graduada (MRG) de Samejima (2010) operando el programa MULTILOG (Thissen, 2003). El MRG utiliza dos pasos para calcular la probabilidad que tiene una persona con nivel de rasgo θ de escoger una categoría *j (j=0, …, m)* en un ítem *i* de *m+1* opciones de respuesta. Primero se definen las probabilidades de optar por la categoría j o una superior según:

Para j=0 se define $P_{i0}^{*}(\theta) = 1$

Para j=1,…, m se define $P_{ij}^{*}(\theta) = \frac{1}{1+e^{-1.7a_i(\theta-b_{ij})}}$

Para j=m+1 se define $P_{i(m+1)}^{*}(\theta) = 0$

Donde a_i es el parámetro de pendiente de cada ítem y b_{ij} (definido únicamente para j=1,…, m) es una serie de parámetros de umbral que separan las categorías adyacentes de la escala de respuesta graduada. Posteriormente, en el segundo paso, se definen las Curvas Características de las Categorías de los ítems a partir de una resta de las probabilidades acumuladas a derecha:

$$P_{ij}(\theta) = P_{ij}^{*}(\theta) - P_{i(j+1)}^{*}(\theta) \; para \; j = 0,\ldots, m$$

La estimación de los parámetros fue por Máxima Verosimilitud Marginal (MML). Durante la fase de calibración se estimaron los parámetros de los nueve ítems correspondientes a cada faceta. Esto implica estimar un parámetro de pendiente (a) y tres parámetros de localización (b_1, b_2, b_3). También se estimaron seis parámetros θ por sujeto, los cuales cuantifican su nivel del rasgo en cada faceta.

El ajuste del MRG se estudió mediante los métodos gráficos y los índices de bondad de ajuste que proporciona MODFIT (Stark, 2001). Este programa permite comparar en 25 niveles del rasgo las probabilidades observadas y esperadas para cada categoría de respuesta al ítem. De esta manera, brinda información para definir si el modelo predice adecuadamente las curvas empíricas. También se obtuvieron los estadísticos χ^2 para cada ítem (9 simples) y para todas las combinaciones de pares y tríos de ítems correspondientes a una misma faceta. Siguiendo a Drasgow, Levine, Tsien, Williams y Mead (1995), se consideraron que los valores de la ratio χ^2/gl superiores a 3 reflejan problemas de desajuste del modelo a los datos. Estos estudios se complementaron con otros indicadores indirectos de ajuste recomendados por Rubio, Aguado, Hontangas y Hernández, 2007): a) una cantidad razonable de iteraciones en la estimación para alcanzar la convergencia, b) parámetros estimados con valores acordes a lo esperable y c) errores de estimación relativamente bajos.

Evidencias de validez y confiabilidad. Si bien se planteó la realización de análisis factoriales para verificar el supuesto de unidimensionalidad de cada faceta, estos estudios son útiles exclusivamente para la aplicación de la TRI. Sin embargo, desde una perspectiva teórica resulta indispensable verificar si la estructura interna del banco completo se corresponde con el modelo jerárquico del Neuroticismo que se establece desde el FFM. Se llevó adelante un AFC de segundo orden considerando la matriz de correlaciones policóricas de los 54 ítems propuestos. También se buscaron evidencias de validez concurrente al estudiar la asociación entre los θ estimados con el MRG y los puntajes de las escalas del EPQ-RS y SCL-90-R.

Para estimar la confiabilidad basada en la consistencia interna se aplicaron los coeficientes alfa de Cronbach y alfa ordinal (Elosua & Zumbo, 2008). Ambos indicadores y sus respectivos intervalos de confianza del 95% (técnica de *boostrap*) se calcularon con la función *scaleReliability* del paquete *userfriendlyscience* (Peters, 2014) del programa R. Desde la TRI se obtuvieron el coeficiente de fiabilidad marginal y la función del error estándar de medida para los valores estimados de θ en cada faceta.

Determinación del algoritmo adaptativo. Se usó el programa Firestar (Choi, 2009) para realizar la simulación post hoc de una administración adaptativa de los ítems de cada faceta a partir de datos reales obtenidos con la muestra de n=352. Este procedimiento consiste en simular una selección progresiva de los ítems que serían presentados en el caso de estar respondiendo un TAI. La implementación se efectuó considerando a la media del rasgo como una estimación inicial razonable del θ del evaluado para dar comienzo a la administración. Luego de recuperar la respuesta a cada ítem, la estimación provisional de θ se efectuó con el método bayesiano de estimación esperada a posteriori (EAP), utilizando a la normal estándar como distribución a priori. En la selección progresiva de los ítems se utilizó el método de máxima información de Fisher, el cual consiste en presentar al evaluado el ítem que resulte más informativo para el nivel provisional de θ estimado hasta ese momento. Finalmente, se decidió adoptar un criterio de parada de longitud variable a fin de garantizar un error mínimo de estimación en el nivel de rasgo igual o menor a .50 (confiabilidad clásica de .75 aproximadamente). No obstante, se especificó una administración mínima de cuatro ítems antes de interrumpir la evaluación para alcanzar una mejor representatividad del contenido.

Para establecer si el algoritmo adaptativo brinda un nivel de rasgo aproximado al que obtendría la persona al responder los nueve ítems de cada faceta (eficiencia), se correlacionaron los niveles θ estimados a partir del TAI con los θ estimados al responder todos los ítems. Además, se reiteraron estudios de validez basados en la estructura interna del constructo (AFC) y de validez concurrente con EPQ-RS y SCL-90-R para considerar si la administración adaptativa altera las propiedades psicométricas.

RESULTADOS

Depuración: Juicio experto y prueba piloto

Sobre la base de los dos criterios definidos para valorar el acuerdo entre jueces (congruencia del ítem con la definición y relevancia) fueron descartados: un ítem de Depresión, dos ítems de Ansiedad, cuatro de Vulnerabilidad y Autoconciencia, cinco de Hostilidad y seis de Impulsividad. A pesar de que 68 ítems fueron valorados positivamente por los jueces, se decidió ensayar igual cantidad de ítems para todas las facetas, lo que redujo a 54 los ítems que pasaron a la fase de calibración. Con respecto a la prueba piloto, no se detectaron dificultades en la comprensión del contendido de los ítems ni la consigna de administración que hicieran necesario realizar cambios.

Calibración de los ítems

Unidimensionalidad de las facetas. La tabla 1 muestra los índices de ajuste obtenidos en los análisis factoriales confirmatorios. Los resultados reflejan un ajuste al modelo unidimensional aceptable (TLI y CFI > .90; RMSEA<.08) para cada una de las facetas del Neuroticismo. En todos los casos las respectivas cargas factoriales para la solución unidimensional fueron estadísticamente significativas (p < .05).

Aplicación del MRG. La tabla 1 resume los resultados conseguidos en el proceso de estimación de los parámetros de los ítems. Se necesitaron entre 20 y 29 iteraciones según la faceta para alcanzar el criterio de convergencia. Los ítems registraron parámetros a con valores entre moderados (para Hostilidad, Impulsividad y Vulnerabilidad) y altos (para Ansiedad, Depresión y Autoconciencia). Con relación a los parámetros de localización b_1, b_2 y b_3, la mayoría se encontraron dentro de un rango acorde a lo esperable (i.e. entre -3 y 3 aproximadamente). No obstante, Hostilidad evidenció mayor amplitud total en los valores de b de los ítems (entre -3.47 y 3.26) y, en general, también se observó una mayor distancia entre los b de cada ítem. Este último resultado, sumado a su bajo valor del parámetro a, configuran curvas ligeramente más aplanadas que en el resto de las facetas y muestran una menor capacidad discriminativa de las categorías de respuesta. También conviene destacar los b de Depresión, dado que adoptaron un rango de actuación comparativamente más acotado que el resto de las facetas.

Tabla 1.
Unidimensionalidad, calibración de ítems y estudios de confiabilidad.

	Ansiedad	Hostilidad	Depresión	Autoconciencia	Impulsividad	Vulnerabilidad
Unidimensionalidad						
CFI	.974	.932	.983	.962	.978	.958
TLI	.966	.909	.978	.949	.971	.944
RMSEA	.066	.070	.054	.073	.046	.056
(90 IC) LI	.054	.058	.042	.061	.034	.044
(90 IC) LS	.078	.082	.066	.085	059	.069
Cargas factoriales						
Mín	.445	.401	.573	.414	.402	.399
Máx	.780	.697	.758	.766	.727	.741
Calibración de ítems con TRI						
Iteraciones	20	20	29	27	24	24
Parámetros a						
Media	1.44	1.07	1.53	1.36	1.20	1.12
Mín (SE)	0.81(.10)	0.68 (.09)	1.17 (.11)	0.76 (.08)	0.74 (.08)	0.62 (.09)
Máx (SE)	2.18(.14)	1.69 (.12)	2.03 (.13)	2.04 (.13)	1.85 (.13)	1.97 (.15)
Parámetros b						
Mín (SE)	-2.42 (.28)	-3.47(.50)	-1.08 (.10)	-0.90 (.18)	-1.80 (.20)	-1.36 (.19)
Máx (SE)	2.35(.28)	3.26 (.41)	2.14 (.21)	3.27 (.42)	2.34 (.23)	3.38 (.42)
Indicadores globales de confiabilidad						
Alfa de Cronbach	.81	.70	.83	.78	.75	.70
[95 IC]	[.79,.83]	[.67,.73]	[.82,.85]	[.76,.81]	[.72,.77]	[.67,.73]
Alfa ordinal	.84	.77	.87	.83	.80	.76
[95 IC]	[.83,.86]	[.75,.80]	[.86,.89]	[.71,.85]	[.78,.82]	[.74,.79]
Fiabilidad Marginal	.84	.75	.84	.82	.78	.76

Los gráficos de ajuste que proporciona MODFIT permitieron verificar que las curvas características de las categorías de respuesta de todos los ítems se mantuvieron dentro del intervalo de confianza asociado a la probabilidad observada de los distintos niveles del rasgo para los que fueron contrastados. Si se consideran los índices de bondad de ajuste, en la tabla 2 se reproducen las distribuciones de frecuencias y resúmenes estadísticos de los valores de las ratios χ^2/gl obtenidos con MODFIT. Como se puede observar, todos los valores de ratio resultaron menores a 3, señalando un ajuste aceptable del modelo a los datos.

Evidencias de validez y confiabilidad

Al analizar los ítems del banco en un estudio factorial confirmatorio de segundo orden, se encontró un ajuste aceptable al modelo jerárquico de seis facetas conforme se postula desde el FFM (*CFI* = .91; *TLI* = .91; *RMSEA* = .045, 90 *IC*[.043, .047]; *SRMR* = .028). En la Figura 1 se puede apreciar que todos los pesos estandarizados de los ítems fueron iguales o mayores a .40, por lo que pueden ser considerados como adecuados (Byrne, 2012).

Todas las facetas correlacionaron de manera moderada con la operacionalización global de la dimensión Neuroticismo que se propone en el EPQ-RS (tabla 4). El estudio de la asociación entre las facetas de Neuroticismo y las escalas del SCL-90-R también registró resultados acordes a lo esperable desde una perspectiva teórica (tabla 4). Las distintas intensidades en las correlaciones reflejan una mayor o menor cercanía de los patrones sintomáticos que mide SCL-90-R y con la red nomológica que puede construirse en torno a cada faceta. En particular, Depresión y Ansiedad son

las que evidenciaron correlaciones más altas con la mayoría de los dominios sintomáticos.

Tabla 2.
Frecuencias y estadísticos descriptivos de la ratio χ^2/gl para evaluar el ajuste al Modelo de Respuesta Graduada.

	Tabla de frecuencias			*M*	*DE*
	< 1	1 < 2	2 < 3		
Ansiedad					
Simple	9	0	0	0.029	0.040
Pares de ítems	18	16	2	1.011	0.577
Tríos de ítems	35	47	2	1.132	0.333
Hostilidad					
Simple	9	0	0	0.011	0.009
Pares de ítems	20	13	3	1.101	0.508
Tríos de ítems	23	61	0	1.223	0.303
Depresión					
Simple	9	0	0	0.045	0.038
Pares de ítems	17	17	2	1.071	0.507
Tríos de ítems	29	54	1	1.122	0.269
Autoconciencia					
Simple	9	0	0	0.024	0.024
Pares de ítems	6	27	3	1.394	0.505
Tríos de ítems	6	76	2	1.378	0.264
Impulsividad					
Simple	9	0	0	0.018	0.016
Pares de ítems	8	20	8	1.517	0.659
Tríos de ítems	1	78	5	1.536	0.306
Vulnerabilidad					
Simple	9	0	0	0.020	0.018
Pares de ítems	7	20	9	1.505	0.571
Tríos de ítems	4	77	3	1.451	0.277

Los índices globales de confiabilidad obtenidos desde la teoría clásica y la TRI fueron adecuados para todas las facetas (tabla 1). La mediana del alfa de Cronbach de las facetas fue de .76 y mejora al considerar la naturaleza ordinal de los datos (*Mdn* alfa ordinal = .81). Aun así, cabe señalar que comparativamente Hostilidad y Vulnerabilidad presentaron coeficientes más bajos que el resto de las facetas. En la Figura 2 se representan los errores estándares de estimación en función del puntaje correspondiente a cada faceta. Todas las funciones adoptan valores de error superiores a .50 (equivalente a una confiabilidad clásica de .75) en el rango de θ entre -2 y 2. Esto implica que el conjunto de ítems que mide cada faceta aporta buena información en un rango considerable de las variables. A excepción de Ansiedad y Hostilidad, los errores de las demás facetas tienden a aumentar hacia los niveles más bajos de los rasgos.

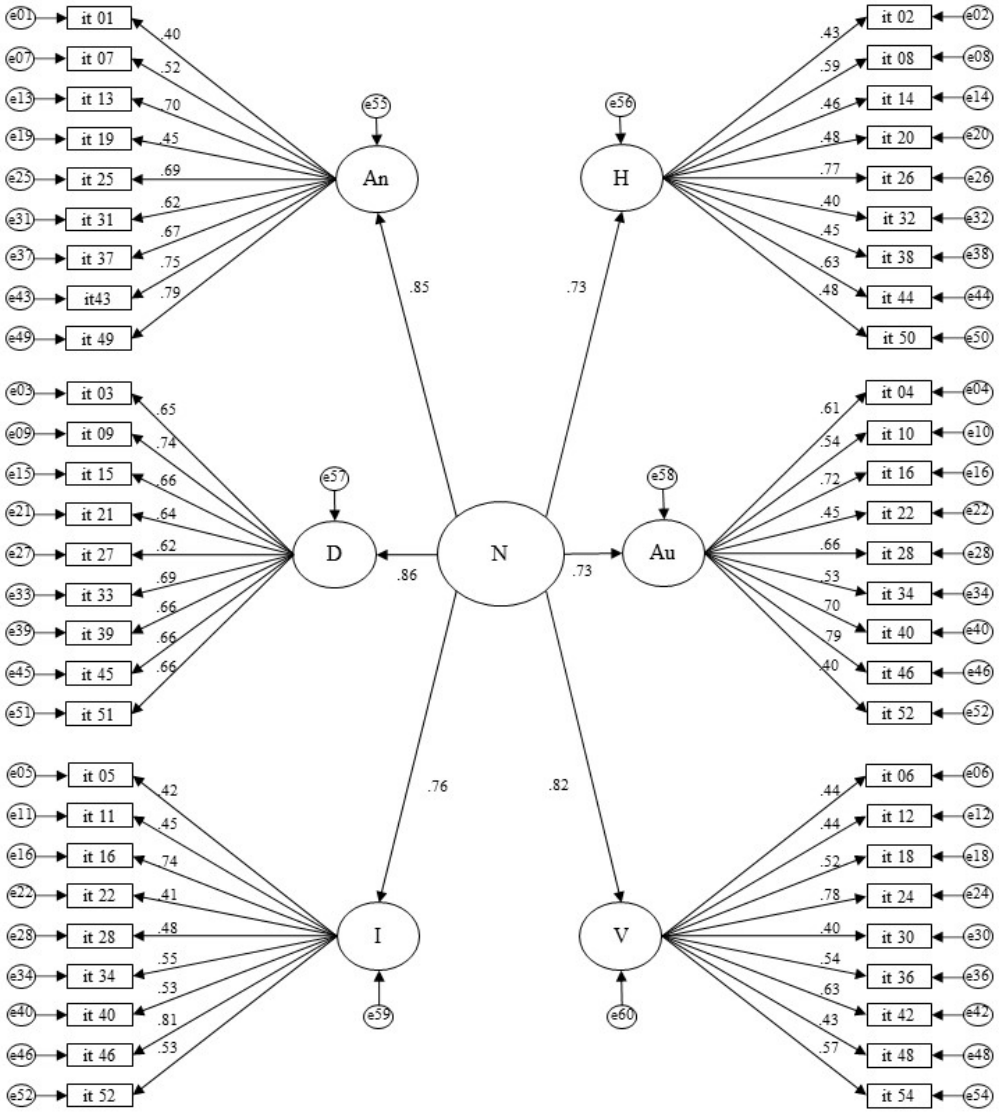

Figura 1. Estructura factorial del banco de ítems de Neuroticismo. N: Neuroticismo. An: Ansiedad. H: Hostilidad. D: Depresión. Au: Autoconciencia. I: Impulsividad. V: Vulnerabilidad

Simulación del TAI

Precisión y eficiencia del TAI. En la tabla 3 se resumen los resultados obtenidos tras la simulación del TAI adoptando un criterio de parada de longitud variable. Se registraron correlaciones altas y positivas (iguales o superiores a .96) entre los valores θ estimados para cada faceta con los nueve ítems y con su respectiva versión adaptativa. La cantidad de ítems promedio que necesitó administrarse para cumplir con el criterio de parada varió de manera notoria entre las facetas. Mientras que Ansiedad demandó la cantidad promedio más baja (4.52 ítems), Hostilidad requirió administrar el promedio más alto (7.96 ítems). Si se compara con relación a los 54 ítems administrados, los sujetos respondieron 36 ítems en promedio, lo que supone una media de 6.07 ítems por faceta. Esta reducción implica una disminución media del 32.6% de los ítems con respecto a la versión completa del instrumento.

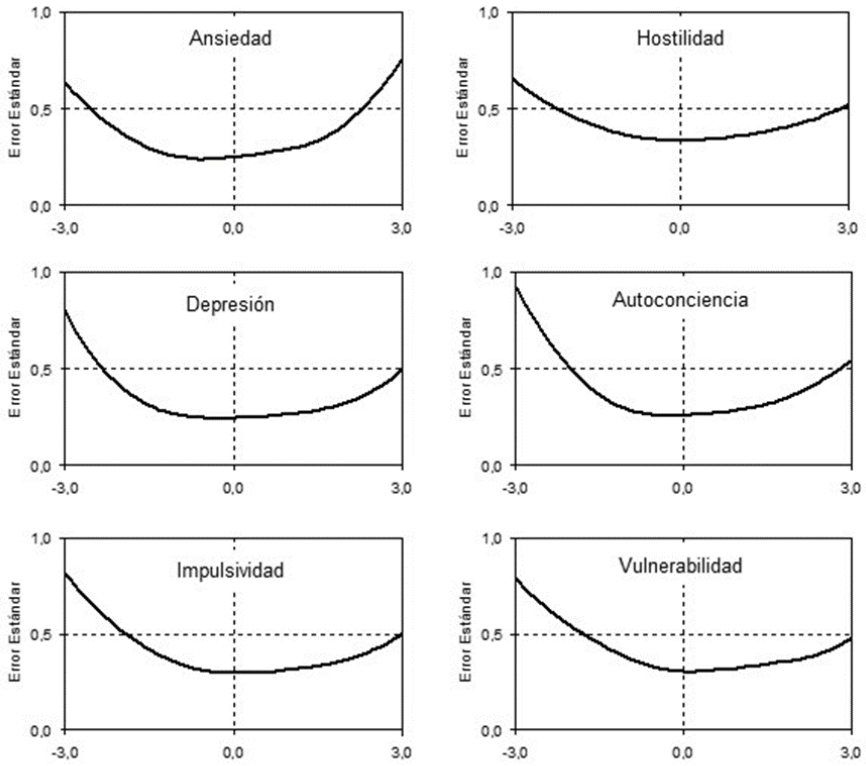

Figura 2. Error estándar en función del nivel de rasgo de cada faceta.

Tabla 3.
Propiedades de la evaluación adaptativa.

	Ansiedad	Hostilidad	Depresión	Autoconciencia	Impulsividad	Vulnerabilidad
Correlación entre TAI y test completo	.97	.99	.96	.96	.98	.99
Cantidad promedio de ítems	4.52	7.96	4.95	5.09	6.81	7.05
SE mediana	0.46	0.49	0.47	0.47	0.49	0.49
Se Máx	0.50	0.60	0.51	0.54	0.57	0.59
Se Mín	0.38	0.47	0.42	0.39	0.45	0.47
% de sujetos con error > .50	0.28	47.7	0.28	9.4	25.3	39.2

Las medianas de los SE de las facetas estuvieron próximas al punto de corte prefijado para la interrupción de la administración (error=0.50). Estos valores oscilaron entre .46 (Ansiedad) y .49 (Hostilidad, Impulsividad y Vulnerabilidad) y se corresponden coeficientes de confiabilidad clásica de entre .79 y .76. El error máximo registrado en la estimación de todos los θ fue de .60 y se observó en el TAI de Hostilidad. Esto significa que ningún individuo fue evaluado con una confiabilidad menor a .64.

Para las facetas Ansiedad, Autoconciencia y Depresión, la mayoría de los evaluados alcanzaron el criterio de finalización antes de administrar los nueve ítems. En cambio, para Impulsividad, Vulnerabilidad y Hostilidad resultaron considerablemente más elevados los porcentajes de individuos para los que no se llegó al error prefijado del θ y se interrumpió la evaluación por lograr el tope de cantidad de ítems (entre 25.3% y 47.7%).

Evidencias de validez del TAI. Se realizó un AFC para verificar si se mantienen las evidencias de validez sobre la estructura interna del Neuroticismo con las estimaciones de θ que ofrece el TAI para cada sujeto. A diferencia de los otros estudios factoriales realizados a nivel del ítem, aquí se debió analizar los puntajes θ de cada faceta porque los sujetos respondieron diferentes subconjuntos de elementos del TAI. Se utilizó la matriz de correlaciones de Pearson asumiendo la continuidad de los puntajes θ. Los parámetros se estimaron con el método máxima verosimilitud robusto (ML Robusto) luego de comprobar que se rechaza la normalidad multivariada de los datos (*Mardia* = 51.5). Los índices de bondad de ajuste calculados (*CFI* = .97; *TLI* = .95; *RMSEA* = .058, 90% *IC* [.027, .091]; *SRMR* = .031) muestran que las seis facetas se agrupan en un dominio general tal como se define desde el FFM. Los coeficientes de regresión oscilaron entre .85 (Ansiedad) y .59 (Autoconciencia). Por otro lado, en la tabla 4 puede cotejarse que el TAI también arrojó puntuaciones que correlacionaron con similar sentido e intensidad con las dimensiones del EPQ-RS y SCL-90-R. Esto significa que la reducción en la cantidad de los ítems que presenta la versión adaptativa no impactó sustantivamente en la estructura interna del constructo ni en su asociación con otras variables.

Tabla 4.
Correlaciones de las estimaciones del total de ítems y del TAI con dimensiones de otros tests.

	Ansiedad		**Hostilidad**		**Depresión**		**Autoconciencia**		**Impulsividad**		**Vulnerabilidad**	
	***θ* Total**	***θ* TAI**	***θ* Total**	***θ* TAI**	***θ* Total**	***θ* TAI**	***θ* Total**	***θ* TAI**	***θ* Total**	***θ* TAI**	***θ* Total**	***θ* TAI**
EPQ-RS												
N	.47**	.45**	.37**	.36**	.47**	.45**	.36**	.39**	.43**	.41**	.36**	.37**
E	.08*	.07	.09*	-.09	.10**	.09	-.28**	-.29**	-.01	.01	.10**	-.10
P	-.03	-.01	-.08*	.09	-.03	.04	.05	-.04	.18**	.17**	-.03	.04
S	-.11**	-.09	-.10**	.09	.12**	.11*	-.06	.08	-.13	.12*	-.13**	-.11*
Escalas del SCL90-R												
So	.32**	.27**	.06	.07	.30**	.30**	.29**	.29**	.14**	.18**	.22**	.18**
Ob	.44**	.40**	.22**	.24**	.50**	.52**	.42**	.42**	.44**	.43**	.46**	.41**
SI	.44**	.43**	.38¨**	.41**	.55**	.63**	.30**	.24**	.27**	.29**	.31**	.24**
Dep	.59**	.56**	.38**	.37**	.69**	.70**	.48**	.46**	.30**	.34**	.52**	.44**
An	.46**	.44**	.11**	.12*	.42**	.40**	.27**	.27**	.15**	.14**	.27**	.23**
Ho	.12**	.14**	.46**	.50**	.25**	.39**	.09*	.08	.38**	.33**	.14**	.10
AF	.47**	.36**	.03	.01	.46**	.39**	.56**	.56**	.13**	.14**	.38**	.32**
IP	.32**	.30**	.31**	.35**	.49**	.57**	.28**	.25**	.40**	.35**	.27**	.22**
Psi	.54**	.47**	.23**	.25**	.60**	.61**	.50**	.48**	.24**	.21**	.44**	.40**
ISG	.55**	.50**	.32**	.35**	.62**	.67**	.45**	.44**	.37**	.37**	.46**	.40**
TSP	.49**	.44**	.29**	.32**	.60**	.66**	.46**	.43**	.39**	.38**	.46**	.41**
IMP	.34**	.35**	.24**	.26**	.23**	.24**	.12**	.15**	.15**	.18**	.12**	.07

Nota. *p < .05; **p < .01. N=Neuroticismo, E=Extraversión, P=Psicoticismo (EPQ-RS), S=Sinceridad, So=Somatizaciones, Ob=Obsesiones y Compulsiones, SI=Sensitividad Interpersonal, Dep=Depresión, An=Ansiedad, Ho=Hostilidad, AF=Ansiedad Fóbica, IP=Ideación Paranoide, Psi=Psicoticismo (SCL-90-R), ISG=índice de Severidad Global, TSP=Total de Sintomas Positivos, IMP=Índice de Malestar Positivo.

COMENTARIOS

En este artículo se compendian los resultados de la primera etapa en la construcción del banco de ítems para la medición de las facetas del Neuroticismo. Otros autores han perseguido como objetivo una evaluación adaptativa realizada a nivel del dominio (e.g. Ferrando, 2001; Rubio et al., 2007), pero la medición a nivel de las facetas supone el beneficio de

discriminar la multiplicidad de emociones y sentimientos negativos que lo constituyen para alcanzar una mayor exhaustividad en la descripción y predicción de los perfiles de los evaluados (e.g. Paunonen, Haddock, Forsterling, & Keinonen, 2003). No obstante, la decisión de priorizar las facetas no invalida la posibilidad de considerar a futuro una medida unidimensional de Neuroticismo a partir de una combinación adaptativa de los ítems del banco si se realizan análisis psicométricos complementarios.

Otro aspecto que amerita ser discutido es la decisión de adoptar la taxonomía de facetas planteadas por McCrae y Costa (2010). La propuesta de estos autores es un muestreo racional artificial e imperfecto de las posibles facetas asociadas al Neuroticismo (Simms, Willams & Simms, 2017). No obstante, posee la ventaja de haber configurado un vocabulario o nomenclatura estándar que favorece la reunión, comunicación y comparación de los resultados obtenidos durante el proceso de validación. En ausencia de un modelo sólido que reconozca variantes émicas, parece más prudente adoptar un enfoque *top-down* que permita elaborar un instrumento fundamentado en una estructura replicada a nivel internacional.

El proceso de diseño y posterior de depuración de los ítems permitió recolectar las evidencias de validez suficientes tanto en el área de contenido como en un área formal (validez aparente). Los estudios factoriales corroboraron la unidimensionalidad de cada faceta (supuesto requerido para la aplicación de TRI). Adicionalmente, se verificó que las mismas se ajustan adecuadamente a la estructura unidimensional de segundo orden que se propone desde el FFM, mostrando evidencias de validez acerca de la estructura interna del test.

Las evidencias de validez basadas en fuentes externas fueron aceptables tanto si se consideran las asociaciones registradas con EPQ-RS como con SCL-90-R. Los puntajes arrojados por cada faceta correlacionaron conforme a lo esperable con la dimensión Neuroticismo del EPQ-RS. Las asociaciones de las facetas con el resto de las dimensiones del EPQ-RS fueron más bajas en la medida en que no se encuentran directamente relacionadas a nivel conceptual. Las correlaciones con las escalas del SCL-90-R también aportan evidencias de validez concurrente. Las facetas del Neuroticismo describen una predisposición a padecer trastornos psicopatológicos (McCrae & Costa, 2010). En este sentido, es coherente que los individuos con mayores niveles de rasgo en las facetas tengan más chances de experimentar mayor intensidad y/o variedad de sintomatología psicológica durante la última semana. Particularmente, las asociaciones moderadas de las facetas Depresión y Ansiedad con las diferentes dimensiones del SCL-90-R pueden ser consideradas como indicativo de una susceptibilidad para percibir los síntomas evaluados por la prueba.

La confiabilidad de las mediciones de las facetas, evaluada a través de los índices alfa de Cronbach, fue buena (*Mdn Alfa* = .76, *Mín* = .70, *Máx* = .83) y se encuentra en el mismo rango, o incluso por encima, de los valores informados habitualmente en la literatura para los inventarios NEO. En la versión más actual, el NEO-PI-3 (McCrae & Costa, 2010), los alfas de las facetas de Neuroticismo oscilan entre .68 y .81 (*Mdn Alfa* = .76). Los resultados empeoran ostensiblemente para las versiones del NEOPIR adaptadas al español. En el estudio de Costa y McCrae (2008), los alfas variaron entre .56 y .76 (*Mdn Alfa* = .64), mientras que en el trabajo de Sanz y García-Vera (2009) se reportaron entre .49 y .81 (*Mdn Alfa* = .70). La adaptación local del

Inventario IPIP-NEO validada en población de universitarios (Cupani, et al., 2014) registró valores de alfa más elevados (*Mdn Alfa* = .725, *Mín* = .64, *Máx* = .88). Sin embargo, es importante señalar que el IPIP incluye una cantidad importante de ítems redundantes que favorecen un aumento artificial del alfa. La comparación de los alfas obtenidos en estos estudios de adaptación con los hallados en la presente investigación sugieren que el muestreo de contenidos seleccionado para los ítems es más consistente para la evaluación de las facetas en la cultura local. A esto debe sumarse que el uso del alfa de Cronbach para analizar la consistencia interna del test aquí construido podría presentar una infraestimación de la confiabilidad de las puntuaciones (Elosua & Zumbo, 2008) porque los ítems se responden con una escala Likert que tiene una categoría menos que las que se emplean en los tests más reconocidos del FFM. Si se analizan los alfas ordinales, los cuales resultan más adecuados para el formato de respuesta, la mediana de la consistencia interna de las facetas aumenta de manera apreciable.

La aplicación del modelo de Samejima de la TRI mostró que los ítems miden con un nivel de precisión parejo para gran parte del espectro de valores que adoptan los rasgos medidos (aproximadamente entre -2 y 2). Para Depresión, Autoconciencia, Impulsividad y Vulnerabilidad se registraron pocos parámetros de umbral localizados en los valores más bajos del rasgo. Por tanto, los ítems de estas facetas resultaron menos precisos en la estimación de los estos niveles bajos. Si bien se trata de un aspecto para mejorar de cara a la incorporación de nuevos ítems al banco, es preciso considerar que esta disposición de los parámetros de umbral podría explicarse, en parte, por las características inherentes de los rasgos medidos. Como señalaron Reise y Waller (2009), los niveles de precisión varían a lo largo de la escala para algunos constructos, principalmente del contexto clínico, que registran variaciones significativas para los puntajes altos, mientras que los puntajes bajos solo reflejan la mera ausencia de manifestaciones del rasgo.

En relación con el algoritmo adaptativo ensayado, en el presente estudio se propuso el uso de un criterio de error prefijado (igual o inferior a .50) para interrumpir la administración adaptativa. Con esta decisión se ha pretendido garantizar un nivel de precisión mínimo en la medida de las facetas sin alargar innecesariamente la administración. Así pues, se ha buscado homogeneizar la precisión con que se miden las facetas, problema que aparece sistemáticamente en los estudios de validación de los inventarios NEO. Como pudo apreciarse párrafos más arriba, el rango de variación que adopta el alfa de Cronbach muestra que hay facetas que son medidas de manera más confiable que otras (McCrae, Kurtz, Yamagata & Terracino, 2011). Estos inventarios presentan una estructura simétrica del número de facetas por dominio y de la cantidad de ítems por faceta. Pero sostener esta organización rígida hace perder de vista los diferentes grados de complejidad teórica de los constructos evaluados (Simms et al., 2017). La faceta Depresión podría requerir de una menor cantidad de ítems porque resulta conceptualmente más acotada o porque sus indicadores son más discriminativos. En cambio, Hostilidad o Impulsividad son lo suficientemente complejas como para demandar más exhaustividad en su evaluación.

La viabilidad de la medición adaptativa de las facetas ha verificado solo para tres de ellas: Depresión, Ansiedad y Autoconciencia. La medición también fue eficiente dado que para la mayoría de los casos se ha precisado una menor cantidad de ítems que los utilizados en los instrumentos convencionales con

longitud fija (como los inventarios NEO). En cambio, el porcentaje de sujetos que no cumplieron con el criterio de parada dejó en evidencia que los ítems que componen las facetas Impulsividad, Vulnerabilidad y Hostilidad no fueron suficientes para alcanzar el error prefijado. Este resultado no debe interpretarse como un defecto propio de la evaluación mediante TAIs sino como una limitación del conjunto de ítems calibrados para estas facetas. La Función de Información total de cada faceta impone una cota a la máxima precisión que puede obtenerse mediante el TAI. Los ítems pertenecientes a estas escalas presentan parámetros de pendiente más modestos, por lo que cada ítem aporta poca información y, por ende, se debe recurrir a una mayor cantidad de elementos para tratar de alcanzar el error prefijado. Como una potencial solución se podría contemplar un criterio más laxo para detener la administración en estas facetas (e.g. error ≥ .55), pero parece más razonable esperar que la incorporación de ítems más discriminativos en las sucesivas etapas resuelva este problema. Aun con estos problemas en la precisión y eficiencia del TAI, las evidencias de validez factorial y concurrente no se vieron afectadas de manera sustantiva por la medición adaptativa.

Las limitaciones del presente estudio son propias de la etapa inicial en la que se encuentra la construcción del banco y que redundan en las dificultades evidenciadas en la administración adaptativa. El análisis de ítems con TRI ha mostrado las fortalezas y debilidades que tiene la versión actual del banco para orientar la futura incorporación de ítems. Resulta necesario refinar la elección de contenidos en la redacción. Para las facetas Hostilidad, Impulsividad y Vulnerabilidad se buscarán indicadores de la cultura local tendientes a propiciar una mayor capacidad discriminativa para todos los niveles de los rasgos. En cambio, con las facetas Depresión y Autoconciencia solo se apuntará a cubrir los niveles bajos. Los ítems de cada faceta que han presentado un funcionamiento óptimo serán utilizados como anclajes en el diseño de la siguiente versión del protocolo para garantizar que los próximos ítems queden calibrados en la misma escala construida en esta etapa.

En cuanto a las limitaciones del estudio a nivel metodológico, en esta etapa no se han efectuado estudios de funcionamiento diferencial de los ítems (DIF), los cuales se proyectan como último control de calidad de los ítems del banco. Aun así, la detección de existencia de DIF en esta fase inicial del banco podría proporcionar información valiosa sobre la naturaleza del constructo (Smith, 2002). Próximos estudios también contemplarán la calibración de los ítems del banco a partir de modelos de la TRI multidimensionales con el fin de ensayar su implementación adaptativa. La experiencia de algunos investigadores ha resultado prometedora, demostrando una ganancia incremental con respecto a los TAIs unidimensionales en la eficiencia de las mediciones (e.g. Makransky et al., 2013; Paap et al., 2017). No obstante, todavía se discuten aspectos vinculados a la determinación del algoritmo adaptativo como el método de selección de ítems (Smits, Paap, & Böhnke, 2018; Tu, Han, Cai & Gao, 2018) y criterios de interrupción (Wang, Chang, & Boughton, 2013; Yao, 2013).

Para concluir, cabe subrayar que la versión actual del banco reúne propiedades psicométricas adecuadas desde el punto de vista de la teoría clásica de tests. Esto significa que la aplicación de los 54 ítems en un formato convencional de lápiz y papel muestra garantías de calidad suficientes como para realizar una valoración de las diferencias individuales en las facetas de Neuroticismo según las definiciones del FFM. En este sentido, constituye un

aporte tecnológico relevante para los profesionales que trabajan en diferentes ámbitos aplicados. El avance en la informatización de las tareas de evaluación psicológica es paulatino pero inexorable. Por esto, se espera que las próximas etapas permitan superar los inconvenientes identificados en la aplicación adaptativa del banco.

REFERENCIAS

Abal, F. J. P., Auné, S. E., Lozzia, G. S. & Attorresi, H. F. (2017). Funcionamiento de la categoría central en ítems de Confianza para la Matemática. *Evaluar*, *17*(2), 18-31.

Abal, F. J. P., Lozzia, G. S., Aguerri, M. E., Galibert, M. S. & Attorresi, H.F. (2010). La escasa aplicación de la Teoría de Respuesta al Ítem en Tests de Ejecución Típica. *Revista Colombiana de Psicología, 19*(1) 111-122.

Baldasaro, R. E., Shanahan, M. J., & Bauer, D. J. (2013). Psychometric Properties of the Mini-IPIP in a Large, Nationally Representative Sample of Young Adults. *Journal of Personality Assessment*, *95*(1), 74–84. https://doi.org/10.1080/00223891.2012.700466

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

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Vecchione, M. (2007). *BFQ: Manuale*. Firenze: OS.

Casullo, M. (2004). Síntomas psicopatológicos en adultos urbanos. *Psicología y Ciencia Social, 6*(1), 49-57.

Choi, S. W. (2009). Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. *Applied Psychological Measurement, 33*(8), 644-645. https://doi.org/10.1177/0146621608329892

Costa, P. T. & McCrae, R. R. (2008). *Inventario de Personalidad NEO Revisado (NEO PI-R). Inventario NEO reducido de Cinco Factores (NEO-FFI)*. Manual. 3ª edición. Madrid: TEA.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874–888. https://doi.org/10.1037/a0027403

Cupani, M., Pilatti, A., Urrizaga, A., Chincolla, A. & Richaud, M. C. (2014). Inventario de personalidad IPIP-NEO: estudios preliminares de adaptación al español en estudiantes argentinos. *Revista Mexicana de Investigación en Psicología*, *6*(1), 55-73.

Derogatis, L. (1994). *SCL-90-R. Symptom Checklist-90-R. Administration, Scoring and Procedures Manual*. Minneapolis: National Computer System.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of Item Candidates. The PROMIS. *Qualitative Item Review. Medical Care, 45*(5), 12-21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Drake, M. M., Morris, D. & Davis, T. J. (2017). Neuroticism's susceptibility to distress: Moderated with mindfulness. *Personality and Individual Differences, 106*, 248-252. https://doi.org/10.1016/j.paid.2016.10.060

Drasgow, F., Levine, M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143-165. https://doi.org/10.1177/014662169501900203

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the to support Army selection and classification decisions* (Tech. Rep. No. 1311). Arlington, VA: U.S. Army Research.

Elosua, P. & Zumbo, B.D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*(4), 896-901.

Eysenck, H. J. & Eysenck, S. B. G. (1994). *Manual of the Eysenck Personality Questionnaire*. California: EdITS/Educational and Industrial Testing Service.

Ferrando, P. J. (2001). The measurement of neuroticism using MMQ, MPI, EPI and EPQ items: a psychometric analysis based on item response theory. *Personality and Individual Differences, 3*0, 641-656. https://doi.org/10.1016/S0191-8869(00)00062-3

Forbey, J. D. & Ben-Porath, Y. S. (2007). Computerized Adaptive Personality Testing: A Review and Illustration With the MMPI-2 Computerized Adaptive Version. *Psychological Assessment, 19*(1), 14 - 24. https://doi.org/10.1037/1040-3590.19.1.14

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34. https://doi.org/10.1037/0003-066X.48.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. En I. Mervielde, I. Deary, F. De Fruyt, y F. Ostendorf (Eds.), *Personality Psychology in Europe*, (Vol. 7, pp. 7-28). Tilburg: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C.,Cloninger, C. R., & Gough, H. C. (2006). The Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality

domains. *Journal of Research in Personality, 37*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Hajek, A., Bock, J. O. & König, H.H. (2017). The role of personality in health care use: Results of a population-based longitudinal study in Germany. *PLoS One, 12*(7):e0181716. https://doi.org/10.1371/journal.pone.0181716

Hengartner, M. P., Kawohl, W., Haker, H., Rössler, W., & Ajdacic-Gross, V. (2016). Big Five personality traits may inform public health policy and preventive medicine: Evidence from a cross-sectional and a prospective longitudinal epidemiologic study in a Swiss community. *Journal of Psychosomatic Research, 84*, 44-51. https://doi.org/10.1016/j.jpsychores.2016.03.012

Jeronimus, B. F., Kotov, R., Riese, H. & Ormel, J. (2016). Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants. *Psychological Medicine, 46*(14), 2883-2906. https://doi.org/10.1017/S0033291716001653

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

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist, 64*, 241–256. https://doi.org/10.1037/a0015309

Lozano, L. M., García-Cueto, E. & Muñiz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology, 4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

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

McCrae, R. R. & Costa, P. T. (2003). *Personality in Adulthood, Second Edition: A Five-Factor Theory Perspective*. New York: Guilford Press. https://doi.org/10.4324/9780203428412

McCrae, R. R. & Costa P. T. (2010). *NEO Inventories professional manual.* Odessa, FL: Psychological Assessment Resources.

McCrae, R. R., Kurtz, J. E., Yamagata, S. & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28-50. https://doi.org/10.1177/1088868310366253

Milojev, P., Osborne, D., Greaves, L. M., Barlow, F. K. & Sibley, C. G. (2013). The Mini-IPIP6: Tiny yet highly stable markers of Big Six personality. *Journal of Research in Personality, 47*, 936–944. https://doi.org/10.1016/j.jrp.2013.09.004

Morizot, J. (2014). Construct Validity of Adolescents' Self-Reported Big Five Personality Traits: Importance of Conceptual Breadth and Initial Validation of a Short Measure. *Assessment, 21*(5), 580-606. https://doi.org/10.1177/1073191114524015

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

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D. & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29*(3), 390-395.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology, 66(*6), 574-583. http://dx.doi.org/10.1037/h0040291

Olea, J. & Ponsoda, V. (2013). *Tests adaptativos informatizados.* Madrid: Ediciones UNED.

Ormel, J., Bastiaansen, A., Riese, H., Bos, E. H., Servaas, M., Ellenbogen, M., … Aleman, A. (2013). The biological and psychological basis of neuroticism: Current status and future directions. *Neuroscience and Biobehavioral Reviews, 37*, 59–72. https://doi.org/10.1016/j.neubiorev.2012.09.004

Paap, M. C. S., Kroeze, K. A., Glas, C. A. W., Terwee, C. B., van der Palen, J. & Veldkamp, B. P. (2017). Measuring Patient-Reported Outcomes Adaptively: Multidimensionality Matters! *Applied Psychological Measurement, 42*(5), 327-342. https://doi.org/10.1177/0146621617733954

Paunonen, S. V., Haddock, G., Forsterling F., & Keinonen M. (2003). Broad versus narrow personality measures and the prediction of behaviour across cultures. *Europe Journal of Personality, 17*, 413–433. https://doi.org/10.1002/per.496

Penfield, R. D. & Giacobbi, P. R. (2004). Applying a Score Confidence Interval to Aiken's Item Content-Relevance Index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225. https://doi.org/10.1207/s15327841mpee0804_3

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16*(2), 56–69.

Reise, S. P. & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364. https://doi.org/10.1177/107319110000700404

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

Reise, S. P. & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychological Medicine, 46*(10), 2025-2039. https://doi.org/10.1017/S0033291716000520

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

Rubio, V. J., Aguado, D., Hontangas, P. M. & Hernández, J. M. (2007). Psychometric Properties of an Emotional Adjustment Measure. An Application of the Graded Response Model. *European Journal of*

Psychological Assessment, 23(1), 39-46. https://doi.org/10.1027/1015-5759.23.1.39
Samejima, F. (2010). Graded Response Model. En W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume 1: Models* (pp. 95-108). Boca Raton: Chapman y Hall/CRC.
Sanchez, R. O. & Ledesma, R.D. (2009). Análisis psicométrico del Inventario de Síntomas Revisado (SCL-90-r) en población clínica. *Revista Argentina de Clínica Psicológica, XVIII,* 265-274.
Sanz, J. & García-Vera, M. P. (2009). Nuevos Baremos para la Adaptación Española del Inventario de Personalidad NEO Revisado (NEO PI-R): Fiabilidad y Datos Normativos en Voluntarios de la Población General. *Clínica y Salud, 20*(2), 131-144.
Sauer-Zavala, S., Wilner, J. & Barlow, D. H. (2017). Addressing neuroticism in psychological treatment. *Personality Disorders: Theory, Research, and Treatment, 8*(3), 191-198. https://doi.org/10.1037/per0000224
Sibley, C. G. (2012). The Mini-IPIP6: Item Response Theory analysis of a short measure of the big-six factors of personality in New Zealand. *New Zealand Journal of Psychology, 41*(3), 21-31.
Simms, L., Williams, T. F. & Simms, E. N. (2017). Assessment of the Five Factor Model. En T. A. Widiger (Ed.) *The Oxford Handbook of the Five Factor Model* (pp. 353-380). New York: Oxford University Press
Smith, L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28*, 754-763. https://doi.org/10.1177/0146167202289005
Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188*(1), 147-155. https://doi.org/10.1016/j.psychres.2010.12.001
Smits, N., Paap, M. C. S., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research, 27*(4), 1055-1063. https://doi.org/10.1007/s11136-018-1821-8
Soto, C. J. & John. O. P. (2017a). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004
Soto, C. J. & John, O. P. (2017b). The Next Big Five Inventory (BFI2): Developing and Assessing a Hierarchical Model With 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power. *Journal of Personality and Social Psychology, 110*(3), 117 - 143. https://doi.org/10.1037/pspp0000096
Squillace, M., Picón Janeiro, J., & Schmidt, V. (2013). Adaptación local del Cuestionario Revisado de Personalidad de Eysenck. *Evaluar*, 13, 19 – 37.
Stark, S. (2001). *MODFIT: A computer program for model-data fit.* Manuscrito no publicado. University of Illinois: Urbana-Champaign.
Tackett, J. L. & Lahey, B. B. (2017). Neuroticism. En T. A. Widiger (Ed). *The Oxford handbook of the five factor model.* New York: Oxford University Press.
Taylor, N. & De Bruin, G.P. (2006). *BTI. Manual of the Basic Traits Inventory*. Johannesburgo, Sudáfrica: JvR
Thissen, D. (2003). *MULTILOG*. Chicago: Scientific Software International.
Tu, D., Han, Y., Cai, Y., & Gao, X. (2018). Item Selection Methods in Multidimensional Computerized Adaptive Testing With Polytomously Scored Items. *Applied Psychological Measurement, 48*(8), 677-694. https://doi.org/10.1177/0146621618762748
Vittengl, J. R. (2017). Who pays the price for high neuroticism? Moderators of longitudinal risks for depression and anxiety. *Psychological Medicine*, 1-12. https://doi.org/10.1017/S0033291717000253
Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99–122. https://doi.org/10.1177/0146621612463422
Watson, D., Nus, E., & Wu, K. D. (2017). Development and Validation of the Faceted Inventory of the Five-Factor Model (FI-FFM). *Assessment,* 1–28.
Widiger, T. A. (2009). Neuroticism. En M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 129-146). New York; Guilford Press.
Widiger, T. A. & Oltmanns, J. R. (2017). Neuroticism is a fundamental domain of personality with enormous public health implications. *World Psychiatry, 16* (2), 144–145. https://doi.org/10.1002/wps.20411
Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*(1), 3–23. https://doi.org/10.1177/0146621612455687
Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*(4), 185-189. https://doi.org/10.1027/1614-0001/a000148

Recibido 21-05-2018 | Aceptado 24-09-2018