PSYKHE
2023, 32(2), 1-15
https://doi.org/10.7764/psykhe.2021.39025

ISSN 0717-0297
www.psykhe.cl

Análisis Psicométrico del Inventario Alemán de Ansiedad ante los Exámenes basado en el Modelo de Respuesta Graduada

Psychometric Analysis of the German Test Anxiety Inventory based on the Graduated Response Model

Luis Rojas-Torres[1], Luis A. Furlan[2] y Guaner Rojas-Rojas[1]

[1] Instituto de Investigaciones Psicológicas, Universidad de Costa Rica
[2] Laboratorio de Evaluación Psicológica y Educativa, Universidad Nacional de Córdoba, Argentina

El presente artículo tiene como objetivo analizar las propiedades psicométricas del Inventario Alemán de Ansiedad ante los exámenes adaptado a Costa Rica (GTAI-CR), con base en el modelo de respuesta graduada. Para este propósito se aplicó el instrumento a 184 personas (101 hombres, 82 mujeres y 1 persona no identificada con las categorías anteriores). Cada una de las cuatro subescalas del GTAI fue evaluada de manera independiente. Se obtuvo que las subescalas de forma global y sus ítems de forma independiente mostraron un ajuste aceptable al modelo. Las curvas características de cada categoría en cada ítem fueron plausibles para grupos representativos de población. Por otro lado, en cada subescala se calculó el rango donde las estimaciones de estas puntuaciones latentes presentaron precisiones aceptables. Finalmente, se presentan recomendaciones para que las escalas de la GTAI-CR puedan mejorar la precisión de las puntuaciones en las que brindan baja información.

Palabras clave: Ansiedad ante los exámenes, Teoría de Respuesta al Ítem, categorías de respuesta, error de medición, precisión.

The aim of this article is to analyze the psychometric properties of the German Test Anxiety Inventory adapted to Costa Rica (GTAI-CR), based on the Graded Response Model. For this purpose, the instrument administered to 184 people (101 men, 82 women and 1 person no identified with the previous categories). Each of the four subscales of the GTAI was evaluated independently. It was found that the subscales globally and their items independently showed an acceptable fit to the model. The characteristic curves of each category in each item were plausible for representative population groups. On the other hand, for each subscale, the range where the estimates of these latent scores presented acceptable accuracies was calculated. Finally, recommendations are presented for the GTAI-CR scales to improve the precision of the scores in which they provide low information.

Key words: Test anxiety, Item Response Theory, responses categories, measurement error, accuracy.

La Ansiedad ante los Exámenes (AE) se define como una emoción aversiva que surge de valorar una situación evaluativa como un evento amenazante, cuyas respuestas emocionales incluyen elementos fisiológicos, psicológicos, comportamentales y fenomenológicos y cuya respuesta emocional más característica es la preocupación constante de fracasar en el examen (Furlan, 2006; Rojas, 2021).

La AE ha mostrado consistentemente una asociación negativa con las calificaciones en los exámenes, en especial, cuando se consideran únicamente las respuestas cognitivas de la emoción (Bonaccio et al., 2012; Chin et al., 2017; Hannon, 2016; Owens et al., 2012; Szafranski et al., 2012). Según la Teoría de la Interferencia (Sarason, 1984; Wine, 1971) esta relación se debe a que una persona que experimenta un alto nivel de AE no puede concentrar su atención en resolver un examen, lo cual le resta recursos cognitivos para el desarrollo adecuado de las tareas evaluadas.

Luis Rojas-Torres https://orcid.org/0000-0002-9085-2703
Luis A- Furlan https://orcid.org/0000-0002-8415-0596
Guaner Rojas-Rojas https://orcid.org/0000-0002-3064-9631

Este estudio recibió apoyo económico de la Universidad de Costa Rica. El artículo es parte de la tesis para optar al grado de Doctor en Educación de la Universidad de Costa Rica. No existe ningún conflicto de intereses que revelar.

La correspondencia relativa a este artículo debe ser dirigida a Luis Rojas-Torres, Instituto de Investigaciones Psicológicas, Universidad de Costa Rica, apartado postal 2060. Email: luismiguel.rojas@ucr.ac.cr

A partir de lo anterior se puede concluir que la AE representa una desventaja para los sujetos que experimentan esta emoción en niveles elevados. Las personas afectadas por la AE obtienen puntuaciones que subestiman su verdadero dominio del tema, lo cual puede traerles consecuencias negativas en el ámbito educativo, como la reprobación de una materia o la no aceptación a un programa educativo de interés. Además, estas consecuencias pueden desencadenar otras situaciones como reducción de la autoestima, procrastinación o deserción.

La medición del constructo AE es relevante debido a que posibilita la detección de personas que pueden experimentar esta emoción en un nivel potencialmente perjudicial. Ahora bien, para lograr esta medición se requiere de un instrumento que posea las propiedades adecuadas para dicha finalidad. Dentro de estos instrumentos destaca el Inventario Alemán de Ansiedad ante los Exámenes (GTAI; Hodapp, 1991).

El GTAI ha sido analizada en muchos estudios desde la Teoría Clásica de los Test (TCT) y, consistentemente ha mostrado un ajuste aceptable a los criterios establecidos por esta teoría (Heredia et al., 2008; Keith et al., 2003; Mowbray et al., 2015; Piemontesi et al., 2012). Debido a esto se considera que el GTAI es un instrumento apropiado para medir la AE. Por otro lado, a pesar de su uso tan extendido en las investigaciones de la AE, no se ha realizado ningún estudio de las propiedades de esta escala desde la Teoría de Respuesta al Ítem (TRI).

En los instrumentos de medición con ítems dicotómicos, los modelos de TRI estiman una curva de probabilidad de acierto para cada ítem. Esta curva es una función de la puntuación latente de los sujetos en el instrumento y un conjunto de parámetros de los ítems. La modelación de la curva supone que conforme se aumenta la puntuación latente en el constructo, la probabilidad de acierto del ítem aumenta (Reckase, 2009).

Por su parte, en instrumentos con ítems politómicos, por lo general, los modelos de TRI estiman en cada ítem una curva de probabilidad para cada categoría de respuesta. De esta manera, en un ítem con cuatro categorías de respuesta se estimarán cuatro curvas de probabilidad. Esta modelación permite estudiar cuál de las categorías de respuesta es la más plausible para cada nivel de la puntuación latente de interés (Reckase, 2009).

En particular, el análisis de las escalas Likert desde la TRI se puede realizar por medio del modelo de respuesta graduada (MRG; Samejima, 1969), el cual es un modelo que considera que las funciones de probabilidad de respuestas deben tener un carácter ordinal, es decir, que para los grupos de puntuaciones latentes más bajas, la primera categoría es la más probable; luego, en un umbral de la puntuación latente, la segunda categoría pasa a ser la más probable y así, sucesivamente, hasta llegar al grupo de puntuaciones latentes más altas, en el cual la última categoría es la más probable. Este supuesto del MRG es esperable para los ítems de las escalas Likert, ya que al aumentar la puntuación latente de las personas se espera que seleccionen las categorías más altas de los reactivos. Este supuesto es evaluado en la TCT con las correlaciones ítem-total, no obstante, en la TCT no se tiene un modelo que permita determinar en qué intervalos del continuo de puntuaciones latentes son más probables dichas categorías.

La importancia del análisis de la probabilidad de respuesta de las categorías es que permiten determinar si estas son plausibles para grupos relevantes de población. Por ejemplo, unas categorías de respuestas que representen frecuencias bajas (nunca, casi nunca) pueden ser poco plausibles en ítem sobre una conducta con alta frecuencia en la población. Este hecho debe ser evaluado desde el punto de vista del aporte informativo del ítem a la escala, ya que es un reactivo dirigido a una conducta variable restringida a puntuaciones latentes altas. Si la escala está cargada de estos ítems se tendría un problema de representación del constructo, ya que no se estarían considerando las conductas cuyas frecuencias bajas son plausibles en la población.

Otro elemento que se puede estudiar con las curvas de probabilidad es el traslape de las categorías. Por ejemplo, un ítem puede tener cinco categorías de respuesta, pero las tres centrales pueden ser tan similares entre ellas que las personas no sepan cuál de ellas elegir, este fenómeno se asocia a un traslape de las curvas. En estos casos, los ítems presentan un problema de contenido, ya que los individuos no logran diferenciar entre una categoría y otra. Con esta situación dilucidada, los constructores de la escala pueden mejorar la definición de las categorías o reducir la cantidad de estas, lo cual podría mejorar la precisión de las respuestas de los sujetos (Rojas, 2021).

Por otro lado, los modelos de TRI son muy relevantes porque permiten determinar los puntos del continuo de la variable latente donde los ítems logran diferenciar mejor los niveles del constructo. Luego, con base en esta información, se puede determinar el error estándar de la habilidad estimada (error estándar de medida)

y generar un intervalo de confianza de acuerdo con las posibilidades del instrumento. Es decir, en los puntos del continuo donde los ítems no diferencien bien las habilidades habrá altos errores estándar de medida, pero en los puntos donde se logren mejores diferenciaciones, habrá menores niveles de error. Esta capacidad de los modelos de TRI es una de las principales ventajas que tienen sobre los modelos de la TCT, ya que estos suponen que el error estándar de medida es el mismo para todos los niveles del constructo (Embretson, 1996; Martínez et al., 2014).

Con base en lo anterior, cuando se realizan análisis desde la TRI se estudia cuáles son los puntos del continuo donde los ítems brindan una información apropiada, para así determinar cuáles intervalos de habilidad restan por ser representados en un instrumento (Hambleton & Jones, 1993). Por ejemplo, una escala con buenas propiedades psicométricas desde la TCT puede diferenciar adecuadamente personas con bajas puntuaciones verdaderas de las personas con puntuaciones verdaderas medias o altas, pero no así, a las personas con puntuaciones verdaderas medias con las que tienen puntuaciones altas.

En función de la importancia que tienen los análisis desde la TRI en el estudio de las escalas, el objetivo de este trabajo es analizar las propiedades psicométricas del GTAI con base en el modelo de respuesta graduada.

Inventario Alemán de Ansiedad ante los Exámenes

El GTAI es una escala elaborada inicialmente en idioma alemán, la cual presenta ítems de cuatro dimensiones de la AE: emocionalidad, preocupación, interferencia y falta de confianza. Las primeras dos dimensiones son las más apoyadas por la literatura (Cassady & Johnson, 2002). La emocionalidad se refiere a las respuestas fisiológicas de la AE, como la sudoración de las manos, la reducción de la respuesta galvánica, la sensación de hiperventilación o aumento del ritmo cardiaco. Por su parte, la preocupación es la dimensión que agrupa las respuestas cognitivas de la AE relacionadas con los pensamientos asociados con la idea de fracasar en el examen.

La propuesta de la dimensión de interferencia considera la manifestación de la AE asociada a no poder concentrarse totalmente durante la toma del examen. Como se mencionó previamente, según la Teoría de la Interferencia, las personas que experimentan un nivel alto de AE presentan dificultades para dirigir su atención al examen. Por otro lado, en el desarrollo de la escala se consideró que era necesario incorporar una dimensión de la AE asociada a la confianza, debido a que una respuesta de la AE es que las personas reduzcan la confianza en ellas mismas, ya que durante la experiencia de la AE las personas tienden a tener pensamientos aversivos sobre ellas mismas, además de los asociados a reprobar la prueba (Hodapp & Benson, 1997).

La versión original del GTAI está conformada por 30 ítems: 8 de emocionalidad, 10 de preocupación, 6 de interferencia y 6 de falta de confianza. Los ítems de la escala son de formato Likert de 4 puntos y van dirigidos a la medición de la AE de tipo rasgo, es decir la propensión a experimentar la AE en las diversas situaciones de evaluación (Zeidner, 1998).

Las propiedades psicométricas del GTAI han sido evaluadas en varios estudios. Por ejemplo, en Keith et al. (2003) se aplicó la escala a 302 estudiantes alemanes. En esta investigación se encontró que las subescalas presentaron alfas de Cronbach entre 0,88 y 0,90, con excepción de la escala de falta de confianza que tuvo un alfa de 0,80. Las correlaciones entre las escalas estuvieron entre 0,42 y 0,48 y las correlaciones ítem-total promedio entre 0,56 y 0,71. Además, se estudió la estructura factorial de la escala en tres tiempos distintos con los mismos sujetos; se obtuvo que la estructura de cuatro factores correlacionados presentaba un ajuste apropiado a los datos (CFI > 0,95; RMSEA < 0,05). Por su parte, Mowbray et al. (2015) realizó un estudio en Australia con 224 estudiantes universitarios. Se observó que las 4 subescalas presentaron alfas de Cronbach entre 0,88 y 0,90. Además, se analizó el ajuste de la escala a un análisis factorial confirmatorio (AFC) de primer orden con cuatro factores y se obtuvo un ajuste moderado (CFI = 0,85, RMSEA = 0,08). En este AFC se observó que todas las cargas factoriales fueron superiores a 0,50, excepto un único ítem.

El GTAI fue adaptado al español para una investigación en Argentina (Piemontesi et al., 2012). En este estudio participaron 781 estudiantes de la Universidad de Córdoba. Los autores también probaron el ajuste de la estructura factorial de 4 factores, pero con la incorporación de una variable latente de segundo orden y el uso de parcelas de ítems. Los índices de ajuste del modelo indicaron un ajuste aceptable a los datos (CFI > 0,95; RMSEA = 0,067). Por otro lado, las escalas presentaron alfas de Cronbach entre 0,87 y 0,89, con excepción de la escala de preocupación, con un alfa de 0,82.

La investigación de Piemontesi et al. (2012) incorporó dos pequeños cambios de la escala original, debido al análisis de la traducción realizada en Heredia et al. (2008). Los cambios fueron: la eliminación de un ítem de la subescala de preocupación y el cambió de un ítem de la escala de interferencia. De hecho, los dos ítems eliminados en la adaptación al español fueron los que presentaron menor calidad psicométrica en la traducción al inglés utilizada en Australia. La traducción de la escala realizada por Piemontesi et al. (2012) fue la base del instrumento utilizado en Costa Rica.

El Modelo de Respuesta Graduada

El modelo de respuesta graduada (MRG) es un modelo politómico de la Teoría de Respuesta al Ítem (TRI), dirigido a ítems con categorías ordinales, como los utilizados en las escalas Likert. Este modelo plantea que en un ítem con K categorías se deben estimar K-1 curvas de probabilidad en función de la habilidad de los sujetos, similares a las curvas características del ítem de la TRI (curvas características de operación, CCO). En la primera curva se modela la probabilidad de obtener una puntuación mayor o igual a la segunda categoría y, así sucesivamente, hasta la curva K-1, en la que se estima la probabilidad de obtener una puntuación mayor o igual a la K-ésima categoría. Por último, el modelo plantea que en cada una de estas curvas se utilizará el mismo parámetro de discriminación.

La representación matemática de la curva característica de operación de la k-ésima categoría, con k en 1, 2,…, K, es:

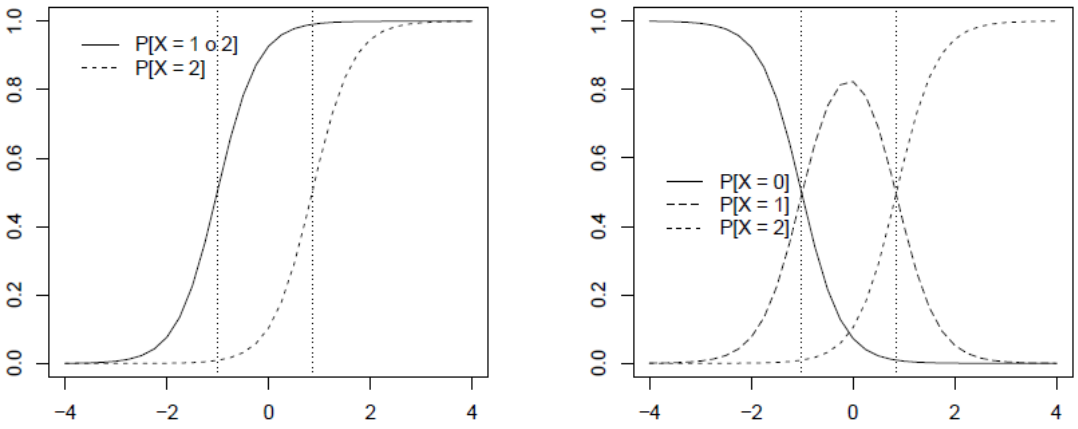
$$P_{ik}^{*}(\theta_p) = P\left(X_{pi} \geq k \,|a_i, b_{ik}, \theta_p\right) = \frac{1}{1 + \exp\left[-1.702\, a_i(\theta_p - b_{ik})\right]}$$

donde X_{pi} es la respuesta del sujeto p al ítem i, θ_p, el parámetro de habilidad de la persona, a_i, el parámetro de la discriminación de la CCO (el cual es igual para las otras CCO del ítem) y b_{ik}, el parámetro de localización de la CCO de la categoría (b_{ik}) (Reckase, 2009). Por otro lado, con las fórmulas de las curvas características de operación se puede calcular la fórmula de la curva característica de una categoría k (CCC), por medio de la resta $P_{ik}(\theta_p) = P_{ik}^{*}(\theta_p) - P_{i(k+1)}^{*}(\theta_p)$ (Reckase, 2009).

A partir de las fórmulas anteriores, se puede deducir que las CCO de las categorías de un ítem tendrán formas logísticas semejantes y solo variarán su ubicación en el continuo de habilidad (esta ubicación la define el parámetro de localización). Por su parte, las CCC tendrán la siguiente forma, la primera será una reflexión de su CCO, las CCC siguientes tendrán formas de montañas y la última, será igual a su CCO. En la figura 1 se presentan las CCO y las CCC de un ítem de la GTAI.

Figura 1

Curvas características del ítem 4 de la GTAI



Nota. A la izquierda las curvas características de operación y a la derecha, las curvas de las categorías individuales. Las categorías 1, 2 y 3 están asociadas a las respuestas 0 = ninguna vez, 1 = algunas veces y 2 = muchas veces, respectivamente.

Método

Participantes

La muestra estuvo conformada por 184 estudiantes de último año de dos secundarias públicas de Costa Rica. La muestra estuvo conformada por 101 hombres, 82 mujeres y una persona que no se identificó en ninguna de las categorías previas. El promedio de edad fue de 17,27 años (0,70). La condición para participar en el estudio fue la realización del examen de admisión a una de las universidades más prestigiosas del país: la Universidad de Costa Rica (UCR), esta condición se debió a que después de este examen se iba a aplicar la GTAI.

Instrumentos

Inventario Alemán de Ansiedad ante los Exámenes adaptado a Costa Rica (GTAI-CR, Rojas, 2021): el GTAI-CR es una escala Likert de tres puntos que mide el nivel de ansiedad ante los exámenes de tipo estado. Esta escala se aplica después de que los sujetos toman un examen de interés, en este estudio se estimó la AE experimentada durante el examen de admisión a la UCR. En la instrucción se les solicita que indiquen la cantidad de veces que experimentaron los enunciados de la escala. Las categorías de respuesta son: 0 = ninguna vez, 1 = varias veces y 2 = muchas veces. Esta escala es una adaptación de la escala homóloga para población argentina (GTAI-AR, Piemontesi et al., 2012). Se divide en cuatro dimensiones: Emocionalidad (8 ítems), Preocupación (9 ítems), Confianza (6 ítems) e Interferencia (6 ítems), las cuales responden a las definiciones establecidas previamente. Por lo general, en el estudio de la AE la escala de Confianza se recodifica para que represente falta de confianza, en este trabajo no se realizó este análisis, debido a que el interés del trabajo es analizar los ítems con base en las respuestas proporcionadas por los sujetos, no en las recodificaciones.

Procedimiento

Para realizar esta investigación se solicitó un espacio en una clase a la que asistían los participantes, con el fin de comunicar el objetivo del estudio y solicitarles colaboración. A las personas que estuvieron de acuerdo con participar en el estudio se les solicitó el consentimiento informado, cuya redacción fue aprobada por el Comité Ético Científico de la Universidad de Costa Rica (UCR). Aproximadamente un mes después, los participantes realizaron el examen de admisión a la UCR. Luego, en una clase que tomaban dos días después del examen de admisión, se aplicó la GTAI-CR.

Análisis de datos

Estudio de simulaciones

El primer paso del trabajo fue un estudio de simulaciones, para determinar si era posible estimar un MRG en un conjunto con 184 observaciones en ocho ítems semejantes a la base de datos de interés. Las 184 observaciones representan el tamaño de muestra con el que se contó en esta investigación y los ocho ítems el tamaño promedio de las subescalas consideradas en la GTAI.

En el análisis preliminar de la TCT de los datos de cada subescala se concluyó que los índices de discriminación eran altos, por lo cual se decidió que en el diseño de simulaciones los parámetros de discriminación también serían altos: entre .8 y 1.2. Con respecto a los parámetros de localización de la CCO de la categoría 1 (b_1) se supuso que provenían de una distribución uniforme y se trabajaron dos condiciones: la ubicación del punto medio de la distribución de estos (-1,5 o -0,5) y b) el radio del rango de variabilidad (0,5 o 1).

Los parámetros de discriminación y localización utilizados en el estudio de simulaciones se generaron por medio de una distribución uniforme definida en los intervalos establecidos. Los parámetros de localización de la CCO de la categoría 2 (b_2) se generaron por medio de la suma del parámetro de localización de la CCO de la categoría 1 más 1 y un valor aleatorio entre 0 y 1. Seguidamente, para cada escenario se generaron 1000 conjuntos de datos con 184 observaciones. Luego, en cada conjunto de datos se estimó el MRG y se contrastaron los coeficientes obtenidos con los parámetros verdaderos.

Análisis de las subescalas

El análisis de las subescalas consistió en la estimación del MRG en cada una de ellas, lo cual implicó el estudio de cuatro salidas del MRG. Antes de estimar dichos modelos, se analizó un AFC de primer orden con cuatro factores determinados por las dimensiones establecidas de la AE en la construcción del instrumento, para verificar que las escalas se configuraban de la forma esperada en la teoría. Para este análisis se utilizó el método de estimación de mínimos cuadrados ponderados por la diagonal y se utilizaron los criterios de que las cargas factoriales fueran superiores a 0,30, que los índices de ajuste CFI y TLI fueran mayores a 0,95, que el índice SRMR fuera menor a 0,08 y que el índice RMSEA fuera menor a 0,06, respectivamente (Cea, 2002; Hu & Bentler, 1999).

En segundo lugar, en cada una de las subescalas de la GTAI se calcularon los indicadores clásicos de la Teoría Clásica de los Test (TCT): medias de puntuación, correlaciones ítem-total y alfas de Cronbach (α), con el fin de contar con un punto de comparación para los resultados obtenidos con la TRI. Se considera que el alfa de Cronbach es aceptable si este supera el umbral de 0,80; similarmente, se considera que una correlación ítem-total es aceptable si esta supera el valor de 0,30. Además, para cada escala se estimó el error estándar de medida utilizado en los intervalos de confianza de las puntuaciones verdaderas en unidades estandarizadas, basado en la aproximación por las puntuaciones empíricas. El error estándar de medida para puntuaciones estandarizadas equivale a $\sqrt{1-\alpha}$ y es igual para cada nivel en la puntuación estimada con la TCT (Muñíz, 2000).

Posteriormente, se estimó el MRG en cada una de las subescalas de la GTAI con el método de máxima verosimilitud. Para analizar la calidad de los ítems se evaluó si los parámetros de los CCO presentaban valores apropiados, a partir de los estándares utilizados en el análisis de las curvas logísticas de los modelos de TRI dicotómicos. El valor de la discriminación de las CCO (a) se considera moderado si es superior a 0,65 y alto si es mayor a 1,35 (Martínez et al., 2014). No obstante, en este estudio, se utilizaron puntos de corte más altos, debido a que el estudio de simulaciones indicó que el MRG con 184 sujetos puede sobreestimar estos valores, por tanto, se utilizaron los puntos de corte de 0,85 y 1,55 (esta decisión se explica en la subsección del estudio de simulaciones en la sección de resultados). Este valor brinda una medida de qué tan pronunciada es la separación de la probabilidad de sujetos con habilidades inferiores a b_{ik} en la escogencia de las categorías superiores o iguales a k, con respecto a la de los sujetos con habilidades superiores a b_k, para cualquiera de los b_k . Cuando un ítem presenta buena calidad psicométrica según la TRI, la separación de las probabilidades es alta, ya que los sujetos con niveles bajos en el constructo deben seleccionar las categorías superiores con baja probabilidad y aquellos con niveles altos las deben seleccionar con alta probabilidad (Ul Hassan & Miller, 2019).

Para analizar los parámetros de localización, se planteó que estos deberían estar dentro de los rangos de puntuaciones que la escala pretende medir, es decir entre -3 y 3, dado que las puntuaciones de los constructos están con unidades estandarizadas. Además, se desea que estos estén separados por al menos una unidad, para evitar la redundancia de las categorías. También se analizó la distribución de estos parámetros, según escala, para determinar si los ítems evaluaban todo el rango de interés o se concentraban en un intervalo específico (Rojas-Torres & Ordóñez, 2019). Para facilitar las interpretaciones de las puntuaciones latentes se plantearon 6 categorías de puntajes: muy bajas, bajas, medias-bajas, medias-altas, altas y muy altas, las cuales correspondieron a los rangos]-∞, -2],]-2, -1],]-1, 0],]0, 1],]1, 2],]2, ∞[, respectivamente.

Para analizar el ajuste de los ítems al modelo se calculó el índice modificado S-χ^2 de Orlando y Thissen (2003), el cual mostró tasas bajas de error tipo I en un estudio de simulación con MRG (Kang & Chen, 2011). Para considerar que un ítem muestra un ajuste aceptable no se debe rechazar la hipótesis nula de que el ítem ajusta al modelo. El ajuste global de los ítems de cada subescala se evaluó por medio de los índices M2 y RMSEA2 (Maydeu-Olivares & Joe, 2006). El índice M2 se evalúa por medio de una prueba estadística, mientras que el RMSEA2 se evalúa por medio de puntos de corte: si es menor que 0,089 se considera adecuado, si es menor que 0,05, bueno y si es menor que $0{,}05/(k-1)$, excelente (Maydeu-Olivares & Joe, 2014). Por último, se procedió a analizar el error estándar de medida, para determinar los niveles del constructo donde las subescalas fueron más precisas. El rango de puntuaciones estimadas con errores estándar de medida menores a 0,39 se denominó intervalo de puntuaciones estimadas con precisión aceptable (el 0,39 es semejante al error estándar obtenido en la TCT para calcular los intervalos de confianza de puntuaciones verdaderas en unidades estandarizadas, cuando el alfa de Cronbach es igual a 0,85).

Todos los análisis se estimaron en el software R en su versión 3.6.1 (R Core Team, 2016). Se utilizaron las librerías mirt (Chalmers, 2012) y lavaan (Rosseel, 2012), para la estimación del MRG y el AFC, respectivamente.

Resultados

Estudio de simulaciones

Los resultados de los cuatro escenarios de simulación indicaron que los coeficientes estimados presentaron un sesgo insignificante con respecto a los parámetros buscados. En cada escenario de simulación, los tres parámetros de los ocho ítems considerados presentaron a lo más un sesgo promedio de 0,04 unidades. Los resultados del estudio de simulaciones se presentan en la tabla 1.

En cuanto a la variabilidad de los estimadores se obtuvo que, en cada escenario de simulación, la mayoría de los estimadores simulados presentaron una desviación estándar de aproximadamente 0,20 unidades. En los parámetros que se observaron mayores variaciones fueron en los de localización b_1 del escenario en el que estos índices eran bajos con alta variabilidad (promedio de las desviaciones estándar de 0,28), esto se debió principalmente a que tres de estos parámetros fueron muy bajos (menores a -2), lo cual lleva a estimaciones menos precisas.

El tamaño de la variabilidad obtenida indica que es esperable que los estimadores de localización posibiliten interpretaciones apropiadas de los parámetros, dado que las desviaciones estándar observadas son pequeñas en comparación con la escala de referencia (estos valores se ubican, generalmente, entre -3,5 y 3,5). Las discriminaciones son los estimadores que deben interpretarse con mayor cautela, dado que la desviación estándar observada es considerable en comparación con el rango común de variación (de 0,2 a 1,5). Por tanto, se concluye que se puede utilizar un MRG en la base estudiada, pero con mayor rigurosidad en el análisis de las discriminaciones, ya que estas pueden estar sobreestimadas. Con base en lo anterior, para este estudio se decidió aumentar los puntos de corte clásicos de las discriminaciones de Martínez et al. (2014), sumándoles una desviación estándar observada en las simulaciones (0,20).

Análisis factorial confirmatorio

Antes del análisis de cada subescala se estimó un AFC con toda la escala utilizando cuatro factores correlacionados, en los que cada factor estaba definido por los ítems de una subescala. Todos los ítems presentaron cargas factoriales superiores a 0,60 en los factores respectivos, con excepción de los ítems 11 y 17 de la dimensión emocionalidad, con cargas de 0,53 y 0,55, respectivamente. Los índices de ajuste presentaron los siguientes valores: RMSEA = 0,038, SRMR = 0,086, TLI = 0,995 y CFI = 0,995. De acuerdo con los criterios de Hu y Bentler (1999) mencionados en la metodología, el modelo presenta un buen ajuste según el RMSEA, el TLI y el CFI; con el SRMR hubo una ligera variación con el punto de corte. De acuerdo con estos señalamientos se aceptó el ajuste del modelo a los datos. Por tanto, la escala presenta la estructura factorial esperada.

Tabla 1

Estadísticos de los estimadores del Modelo de Respuesta Graduada en cuatro escenarios de simulación de una escala de 8 ítems de 3 categorías con 184 personas

	c = -1,5 , r = 0,5			**c = -1,5 , r = 1**			**c = -0,5 , r = 0,5**			**c = -0,5 , r = 1**		
	Par	**Ses**	**DE**	**Par**	**Ses**	**DE**	**Par**	**Ses**	**DE**	**Par**	**Ses**	**DE**
Discriminación de las CCO (a)												
ítem 1	0,81	0,02	0,14	1,19	0,03	0,26	1,03	0,03	0,20	1,14	0,02	0,21
ítem 2	1,10	0,03	0,20	0,89	0,02	0,16	0,82	0,01	0,15	1,01	0,02	0,17
ítem 3	1,16	0,04	0,20	0,94	0,04	0,22	0,90	0,02	0,17	0,94	0,02	0,18
ítem 4	1,13	0,03	0,19	1,00	0,04	0,21	1,13	0,03	0,20	0,86	0,02	0,16

(continúa)

Tabla 1 (Conclusión)

Estadísticos de los estimadores del Modelo de Respuesta Graduada en cuatro escenarios de simulación de una escala de 8 ítems de 3 categorías con 184 personas

	c = -1,5 , r = 0,5			c = -1,5 , r = 1			c = -0,5 , r = 0,5			c = -0,5 , r = 1		
	Par	**Ses**	**DE**	**Par**	**Ses**	**DE**	**Par**	**Ses**	**DE**	**Par**	**Ses**	**DE**
Discriminación de las CCO (a)												
ítem 5	1,03	0,04	0,21	0,89	0,02	0,17	1,03	0,01	0,18	0,82	0,02	0,16
ítem 6	1,18	0,03	0,21	0,97	0,02	0,19	1,07	0,02	0,20	0,92	0,02	0,17
ítem 7	1,02	0,03	0,19	1,05	0,02	0,19	0,96	0,02	0,17	0,95	0,02	0,17
ítem 8	0,88	0,02	0,17	1,05	0,03	0,20	0,87	0,01	0,15	1,19	0,03	0,21
Prom	1,04	0,03	0,19	1,00	0,03	0,20	0,98	0,02	0,18	0,98	0,02	0,18
Localización de la CCO 1 (b_1)												
ítem 1	-1,72	-0,02	0,29	-1,62	-0,04	0,24	-0,97	-0,01	0,17	-1,43	-0,03	0,22
ítem 2	-1,29	-0,01	0,20	-0,93	-0,01	0,19	-0,70	-0,01	0,17	-1,24	-0,03	0,21
ítem 3	-1,20	0,00	0,19	-2,34	-0,05	0,42	-0,48	-0,01	0,15	-1,16	-0,03	0,21
ítem 4	-1,23	-0,01	0,19	-2,45	-0,07	0,42	-0,75	-0,01	0,15	-1,48	-0,03	0,26
ítem 5	-1,94	-0,02	0,30	-1,76	-0,03	0,30	-0,94	-0,02	0,17	0,30	0,00	0,16
ítem 6	-1,23	-0,01	0,18	-2,05	-0,05	0,34	-0,05	0,00	0,13	-1,10	-0,02	0,21
ítem 7	-1,71	-0,02	0,28	-1,44	-0,02	0,22	-0,47	-0,01	0,14	-0,36	-0,01	0,15
ítem 8	-1,98	-0,02	0,32	-0,63	-0,01	0,15	-0,69	-0,01	0,16	-1,00	-0,01	0,17
Prom	-1,54	0,02	0,24	-1,65	0,04	0,28	-0,63	0,01	0,15	-0,93	0,02	0,20
Localización de la CCO 2 (b_2)												
ítem 1	0,09	0,01	0,15	-0,60	-0,01	0,14	0,12	0,00	0,13	0,06	0,00	0,13
ítem 2	-0,22	0,00	0,13	0,28	0,01	0,15	1,24	0,04	0,24	0,59	0,00	0,15
ítem 3	0,57	0,01	0,13	-1,12	-0,02	0,20	1,10	0,02	0,21	-0,07	0,00	0,14
ítem 4	0,55	0,01	0,14	-0,81	-0,01	0,17	0,95	0,01	0,16	-0,19	-0,01	0,15
ítem 5	-0,63	0,00	0,14	0,09	0,00	0,15	0,67	0,01	0,15	1,74	0,05	0,31
ítem 6	-0,19	0,01	0,13	-0,41	-0,01	0,14	1,55	0,04	0,24	0,21	0,00	0,15
ítem 7	-0,44	0,00	0,14	0,37	0,01	0,13	0,97	0,01	0,18	1,15	0,02	0,21
ítem 8	-0,30	0,00	0,14	1,24	0,02	0,20	0,82	0,01	0,18	0,30	0,00	0,13
Prom	-0,07	0,00	0,14	-0,12	0,01	0,16	0,93	0,02	0,19	0,47	0,01	0,17

Nota. c,r = centro y radio del intervalo en el que se generaron los parámetros de localización b1, el c = -1,5 representa valores bajos y el c = -0,5, medios; el r = 0,5 representa variabilidad baja y el r = 1, alta. Par=Parámetro utilizado en la simulación, ses = promedio de los sesgos de los estimadores encontrados con el parámetro original, DE = desviación estándar de los estimadores simulados, Prom = promedio de la columna respectiva; en la columna de sesgo corresponde al promedio de los valores absolutos.

Teoría Clásica de los Test

En la tabla 2 se presentan los estadísticos de los ítems de cada una de las subescalas, desde la TCT. Tomando en cuenta que las puntuaciones de los ítems pertenecen al intervalo 0 a 2, se concluye que los ítems de preocupación son los que ocurren con mayor frecuencia, ya que los promedios de respuesta son mayores a 1 en todos los ítems. En cambio, en las otras escalas, solo dos ítems superan el valor de 1 y por unas pocas centésimas.

Tabla 2

Estadísticos del análisis de las subescalas de la GTAI según la Teoría Clásica de los Test (TCT) y el Modelo de Respuesta Graduada (MRG)

Ítem	TCT		MRG			Test S-χ^2		
	Dif	r IT	a	b1	b2	S-χ^2	gl	p
Preocupación								
2- Pensé en la importancia que el examen tenía para mí	1,72	0,64	1,05	-2,81	-0,94	9,58	10	0,48
5- Me preocupó saber si podía afrontar el examen	1,07	0,75	1,17	-0,90	0,61	12,27	13	0,51
8- Pensé en las consecuencias de fracasar en el examen	1,34	0,83	1,90	-1,16	0,00	8,68	10	0,56
9- Me pregunté si mi rendimiento sería lo suficientemente bueno	1,26	0,67	0,91	-1,54	0,33	7,51	16	0,96
13- Pensé en lo mucho que me importaba obtener un buen resultado	1,67	0,69	1,28	-2,12	-0,81	7,09	10	0,72
16- Me preocupó el resultado de mi examen	1,53	0,81	1,68	-1,45	-0,40	11,25	10	0,34
20- Me preocupó cómo se vería mi calificación	1,39	0,81	1,69	-1,14	-0,20	12,11	12	0,44
22- Me preocupó que algo saliera mal	1,18	0,78	1,37	-1,09	0,42	9,89	11	0,54
26- Pensé en lo que pasaría si me iba mal	1,23	0,79	1,69	-0,92	0,17	5,68	9	0,77
Interferencia								
4- Me bloqueé por los pensamientos que me pasaban por la cabeza	1,03	0,69	0,87	-1,16	0,99	5,97	11	0,88
10- Me distraje por pensar en cualquier cosa	0,91	0,82	1,79	-0,43	0,71	6,40	7	0,49
14- Perdí el hilo de mis pensamientos fácilmente	0,77	0,80	1,47	-0,24	1,07	9,45	8	0,31
18- Tuve dificultades para recordar las cosas debido a que estaba pensando en mis problemas	0,50	0,69	1,07	0,41	1,65	19,77	10	0,03
23- Interrumpí mi razonamiento porque algo de poca importancia llamó mi atención	0,74	0,81	1,80	-0,19	1,06	0,80	7	1,00
29- Tuve dificultades en la concentración debido a que me distraía con algún pensamiento	0,87	0,88	3,63	-0,41	0,81	3,96	5	0,56
Confianza								
1- Tuve seguridad en mi capacidad	1,42	0,63	0,83	-2,52	0,01	5,47	4	0,24
7- Tuve confianza en mi propio desempeño	1,48	0,79	1,84	-2,00	-0,09	3,53	3	0,32
12- Sentí que podía confiar en mí mismo	1,53	0,82	2,64	-1,94	-0,19	2,67	3	0,44
19- Me sentí conforme conmigo mismo/a	1,25	0,73	1,12	-1,66	0,49	5,09	4	0,28
25- Confié que lograría resolver todo el examen	1,53	0,73	1,28	-2,16	-0,23	0,97	4	0,91
28- Me sentí convencido/a de que haría bien el examen	1,29	0,78	1,53	-1,61	0,35	4,54	4	0,34
Emocionalidad								
3- Tuve una sensación rara en mi estomago	0,89	0,74	1,36	-0,51	0,93	5,73	12	0,93
6- Sentí mi cuerpo tenso	0,90	0,69	0,90	-0,41	0,90	9,37	15	0,86
11- Me sentí intranquilo/a	0,82	0,53	0,57	-0,67	1,89	29,13	19	0,06
15- Sentí que mi corazón latía fuertemente	0,50	0,68	1,36	0,40	1,35	16,52	12	0,17
17- Me sentí incómodo/a	0,54	0,52	0,54	0,44	2,43	20,49	19	0,37
21- Temblé de nerviosismo	0,37	0,69	1,69	0,73	1,47	9,83	10	0,46
24- Tuve una sensación de angustia	0,73	0,76	1,23	-0,12	1,11	13,25	13	0,43
27- Me sentí nervioso/a	1,00	0,77	1,64	-0,64	0,65	16,88	11	0,11

Nota. Dif = Dificultad TCT; r IT. = correlación ítem-total; a = discriminación de las curvas de categorías agrupadas; bi1 = parámetro de localización de la agrupación de las categorías algunas veces y muchas veces; bi2 = parámetro de localización de la respuesta muchas veces; S-χ^2 = estadístico chi cuadrado del test S-χ^2 de ajuste de los ítems al MRG; gl = grados de libertad de la prueba S-χ^2; p = valor p de la prueba S-χ^2.

Por otro lado, los ítems presentaron correlaciones ítem-total muy altas, todas superiores a 0,50, lo cual indica que los ítems efectivamente están asociados con las puntuaciones de la escala respectiva. Las correlaciones ítem-total promedio de cada subescala fueron 0,75, 0,78, 0,75 y 0,67 para Preocupación, Interferencia, Confianza y Emocionalidad, respectivamente.

Luego, el alfa de Cronbach de las escalas fue 0,904, 0,873, 0,841 y 0,827 para Preocupación, Interferencia, Confianza y Emocionalidad, respectivamente. Estos valores indicaron que cada una de las escalas presentó una consistencia interna adecuada. Por otro lado, los errores estándar de medición de las puntuaciones verdaderas, en unidades estandarizadas, fueron 0,31, 0,36, 0,40 y 0,42 para Preocupación, Interferencia, Confianza y Emocionalidad.

Modelo de Respuesta Graduada

En la tabla 2 se muestran los resultados del análisis con el MRG. En cada una de las subescalas todos los ítems mostraron un ajuste apropiado al MRG, ya que todas las pruebas asociadas al estadístico S-χ^2 fueron no significativas. Únicamente el ítem 18 ubicado en la escala Interferencia presentó un estadístico significativo. Dado este resultado, se decidió que este ítem no fuera considerado para la estimación de la puntuación latente de Interferencia. Al revisar este ítem se observó que las personas con puntuaciones latentes altas en Interferencia presentaron un alto porcentaje de selección de la categoría algunas veces, en comparación con la categoría muchas veces, por lo cual, se hipotetiza que esta es la razón por la que el ítem no se ajusta al MRG.

Con respecto al ajuste global del modelo a los datos se obtuvo un ajuste adecuado en todas las escalas, ya que sus RMSEA2 fueron inferiores a 0,089. Los RMSEA2 fueron 0,072, 0,065, 0,052 y 0,053 para Preocupación, Interferencia, Confianza y Emocionalidad, respectivamente (Maydeu-Olivares & Joe, 2014). Además, la prueba estadística del M2 resultó no significativa para todas las escalas, menos Preocupación (Preocupación: M2(18)=34,22, p=0,012; Interferencia: M2(3)=5,29, p=0,152; Confianza: M2(3)=4,50, p=0,212; Emocionalidad: M2(12)=18,03, p=0,115); por lo cual en aquellas escalas no se rechazó la hipótesis de que el modelo se ajusta a los datos. En cuanto a la escala de Preocupación no se rechaza el ajuste del modelo debido a que la prueba estadística del M2 puede ser muy severa, por eso se acompaña con el RMSEA2 para concluir el rechazo (Auné et al., 2020).

Por su parte, en todos los ítems de cada subescala, los índices de localización se ubicaron en el rango de -3 a 3 y, además, los dos parámetros de localización de cada ítem estuvieron separados por al menos una unidad de distancia.

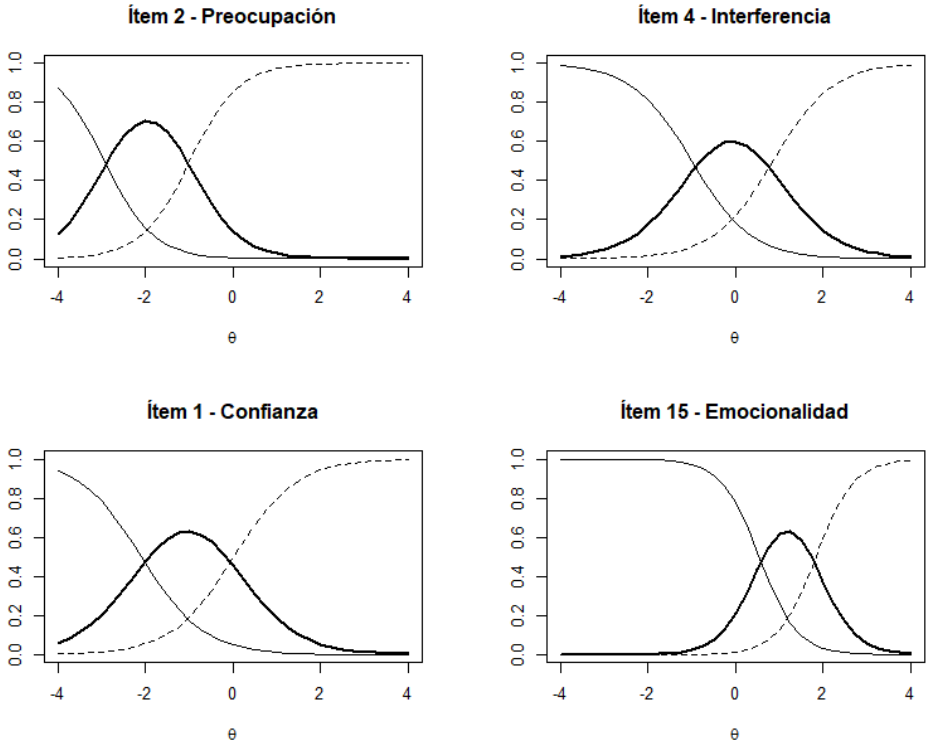
En la escala de Preocupación los parámetros de localización estuvieron concentrados en los valores negativos de la escala. El promedio de los parámetros b_1 fue -1,46 (rango de -2,81 a -0,90), mientras que el promedio de los parámetros b_2 fue -0,09 (rango de -0,94 a 0,61). Ahora bien, con base en las curvas características de las categorías extremas, se sabe que en los valores menores que b_1, la categoría más probable es la 1 (ninguna vez) y que en los valores mayores que b_2, la categoría más probable es la 3 (muchas veces). Por tanto, se concluye que la categoría ninguna vez es plausible solo para puntuaciones muy bajas y la categoría muchas veces para puntuaciones positivas (la mitad del rango de variación de las puntuaciones).

Por su parte, en la escala de Interferencia, los parámetros b_1 y b_2 estuvieron más distribuidos en el intervalo de -3 a 3, el promedio de b_1 fue -0,34 (rango de -1,16 a 0,41) y el de b_2, 1,05 (rango de 0,71 a 1,65). Con respecto a la escala de Confianza, el comportamiento fue similar al de Preocupación, el promedio de b_1 fue -1,98 (rango de -2,52 a -1,61) y el de b_2 fue 0,06 (rango de -0,23 a 0,49). En la escala de Emocionalidad el comportamiento fue inverso al de las escalas de Preocupación y Confianza, ya que promedio de los parámetros b_1 fue -0,10 (rango de -0,67 a 0,73), mientras que el promedio de los b_2 fue 1,34 (rango de 0,65 a 2,43). Por tanto, en la escala de Emocionalidad, la categoría ninguna vez fue muy plausible para las puntuaciones negativas (la mitad del rango de variación de las puntuaciones), mientras que la categoría muchas veces solo para las personas con puntuaciones muy altas.

En la figura 2 se muestra un ejemplo de las CCC de un ítem de cada subescala. Se puede observar que, en todos los ítems, la curva de la categoría central es la más probable para al menos un intervalo de puntuaciones de una unidad de longitud. Por otro lado, en los ítems de Preocupación y Confianza, las curvas características de la categoría muchas veces indican que esta es muy probable para las puntuaciones positivas (las cuales representan casi la mitad de la población); en cambio en el ítem de Emocionalidad se observa que la curva de la categoría ninguna vez es muy probable para las personas con puntuaciones negativas. Por otro lado, en el ítem de Preocupación se puede observar como la categoría ninguna vez solo es plausible para puntuaciones menores que -2.

Figura 2

Curvas Características de Categorías de ítems de la GTAI



Nota. La línea sólida simple es la curva característica de la categoría ninguna vez, la sólida resaltada es la de la categoría algunas veces y la intermitente es la de la categoría muchas veces.

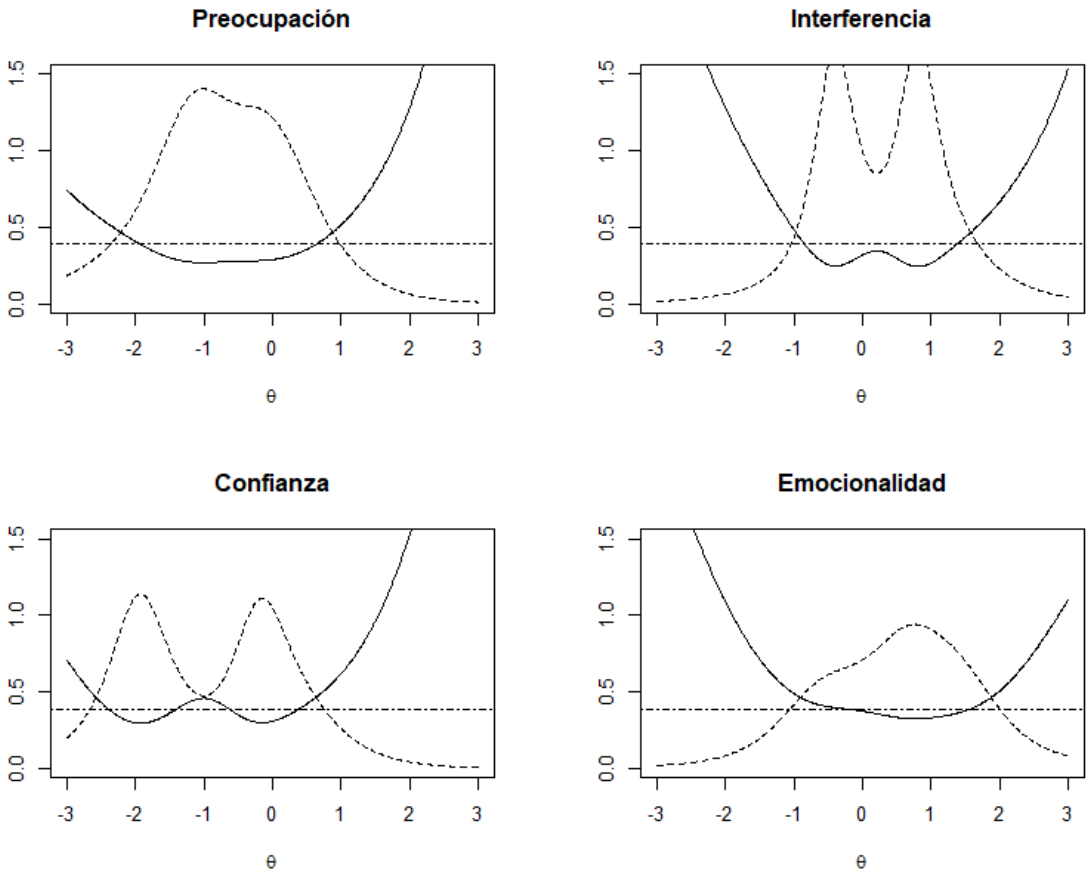
El índice de discriminación de las CCO fue adecuado en la mayoría de los ítems. En todos los ítems los valores fueron superiores a 0,85, con excepción de los ítems 11 y 17 de Emocionalidad que presentaron índices de 0,57 y 0,54 (Martínez et al., 2014). Estos ítems también fueron los que obtuvieron las cargas factoriales más bajas en el AFC inicial y las correlaciones ítem-total más bajas en el análisis con la TCT. A partir del análisis de los índices de discriminación en las curvas características de las categorías extremas, se concluye que, en la mayoría de los ítems, en los puntos de corte b_1 la probabilidad de seleccionar la categoría ninguna vez decrece aceleradamente y en los puntos de corte b_2 la probabilidad de seleccionar la categoría muchas veces aumenta considerablemente.

Por último, en la figura 3 se muestran las curvas de los errores estándar de medición del MRG en cada subescala de AE. Los intervalos de puntuaciones estimadas con precisión aceptable (con un error estándar menor o igual que 0,39) fueron [-1,94, 0,64], [-0,85, 1,38], [-2,40, -1,40] ∪ [-0,63, 0,39] y [-0,25, 1,57] para Preocupación, Interferencia, Confianza y Emocionalidad, respectivamente. El error estándar de medida mínimo que alcanzaron estas escalas fue 0,27, 0,24, 0,30 y 0,33, respectivamente, según el orden anterior. Las escalas de Interferencia y Confianza mostraron errores estándar de medidas altos a distancias cortas por la izquierda de los extremos inferiores de los intervalos señalados, mientras que las escalas de Emocionalidad y Preocupación mostraron errores estándar de medida más moderados en esta dirección. En distancias cortas por la derecha de los extremos superiores de los intervalos, las puntuaciones de todas las subescalas mostraron errores estándar de medida altos.

En este punto se puede ejemplificar la mayor ventaja de los modelos de TRI con respecto a la TCT: la precisión de la medida. Por ejemplo, a las personas con un puntuación TRI de Preocupación igual a 0 y 2 unidades, se les reportará un intervalo de confianza del 95% de su puntuación de 0 ± 0,56 y 2 ± 2,50, respectivamente (los errores estándar de medida para estas puntuaciones fueron 0,29 y 1,28, respectivamente y el radio del intervalo se obtuvo por el producto con 1,96); en cambio, en las puntuaciones estandarizadas TCT de 0 y 2 unidades (que no son análogas a las de TRI) se les reportará un intervalo de

confianza de sus puntuaciones de 0±0,61 y 2±0,61, respectivamente. Lo anterior evidencia que en la TCT se ignora que las personas que obtuvieron puntuaciones altas realmente obtuvieron estimaciones muy imprecisas. En la TRI se muestra que hay problemas en la precisión en las puntuaciones altas de la escala y se observa que una persona con una puntuación alta (como de 2 unidades), podría realmente tener una puntuación media o una extremadamente alta.

Figura 3
Errores estándar de medida y función de información de las subescalas de la GTAI



Nota. La línea sólida es la curva del error estándar de medida y la intermitente la de la función de información (la escala vertical de esta curva es 10 veces la indicada en el eje y).

Discusión

En primer lugar, hay que destacar que el artículo presenta una metodología para el análisis de escalas con base en el Modelo de Respuesta Graduada. En varios estudios con el MRG se utiliza el modelo para estudiar el ajuste al conjunto de datos y describir las características de los ítems. En cambio, en este trabajo se utilizan los parámetros estimados del MRG para determinar si los ítems presentan propiedades deseables de medición: discriminación de las CCO aceptables y separación aceptable de sus localizaciones.

Aunado a lo anterior, se utilizó una propiedad de las curvas características de las categorías extremas para el análisis de los ítems: los intervalos donde estas categorías son las más probables (de hecho, altamente probables, porque la probabilidad de selección se acerca a 1 conforme se avanza en la dirección extrema y la aceleración del crecimiento es determinada por la discriminación). A partir de estas CCC se pudo determinar si alguna categoría extrema era plausible únicamente para un grupo poco representativo o si era muy

plausible para grupos mayoritarios de la población. Luego, si la mayoría de los ítems presentan una de esas características, se estaría ante un problema de subrepresentación del constructo en la escala.

Además, se utiliza un estudio de simulaciones para justificar el uso del MRG en una muestra relativamente pequeña. Esta propuesta es una alternativa para argumentar por qué el uso de un modelo es apropiado para analizar un conjunto de datos y cuáles son las limitaciones de dicho modelo en el estudio de la base de interés. En este trabajo se concluyó que la estimación de las discriminaciones de las CCO era la más afectada por el tamaño de muestra, debido a lo anterior se decidió aumentar el estándar establecido para los estimadores de este parámetro. Por otro lado, es importante mencionar que en el diseño de simulaciones se utilizaron las condiciones que se consideraron más relevantes; en otro estudio se podrían utilizar más condiciones para tener conclusiones generalizables a conjuntos de datos no tan específicos.

En segundo lugar, se discute el análisis de la GTAI con base en la metodología descrita previamente. El estudio mostró que, en cada subescala, el modelo se ajustó apropiadamente a los ítems, los índices de discriminación de las CCO fueron aceptables y los índices de localización de las CCO tuvieron una separación aceptable, lo cual indica que, en cada ítem, todas las categorías se asociaron a un intervalo relevante de las puntuaciones.

La característica de los parámetros estimados que más aportó para el planteamiento de mejoras en la GTAI fue la plausibilidad de las categorías extremas (ninguna vez y muchas veces). En los ítems de las subescalas de Preocupación y Confianza la categoría muchas veces presentó probabilidades muy altas de selección para todas las personas con puntuaciones positivas. Por su parte, en los ítems de la subescala de Emocionalidad se presentó lo contrario, la categoría ninguna vez fue muy probable para personas con puntuaciones latentes bajas. Lo anterior indicó que en las primeras dos escalas los ítems no tenían categorías que permitieran discriminar entre personas con puntuaciones un poco altas de personas con puntuaciones más altas. Estos resultados coincidieron con lo observado en la precisión de las puntuaciones proporcionadas por estas subescalas. Las puntuaciones de Preocupación y Confianza fueron imprecisas en los valores altos, mientras que las de Emocionalidad fueron en los valores bajos.

Por otro lado, la escala que presentó mejores precisiones en las estimaciones de las puntuaciones ubicadas entre -1 y 1 fue la de Interferencia. A diferencia de las otras tres escalas, el intervalo de puntuaciones con precisión aceptable cubrió casi todo el rango de puntuaciones medias-bajas y medias-altas. Estas dos categorías de puntuaciones deben ser estimadas con bajos niveles de error, debido a que ellas generalmente abarcan más de la mitad de la población, ya que en los modelos de TRI, sin equiparación, se estiman las puntuaciones procurando una distribución normal estándar (Baker & Kim, 2004).

En tercer lugar, se presentan las recomendaciones para mejorar las subescalas. Para mejorar la subescala de Preocupación se pueden construir ítems dirigidos a discriminar las personas con niveles de preocupación altos de aquellas con niveles muy altos. Las personas con valores muy altos en el constructo pueden presentar preocupaciones más irracionales, que no serían plausibles en un porcentaje importante de las personas que marcó la respuesta muchas veces en la mayoría de reactivos (Furlan, 2006). Dentro de las preocupaciones irracionales se encuentran los pensamientos catastrofistas que no se presentan en la GTAI, por ejemplo, un ítem semejante a "durante el examen pensé que estaba respondiendo mal todas las preguntas".

De forma similar, en la escala de Confianza hay que mejorar la precisión de las puntuaciones en los niveles medios-altos y altos. Al igual que en la escala de Preocupación, una recomendación sería construir reactivos en los que la categoría muchas veces sea plausible para niveles muy altos, pero no para niveles altos. Estos ítems deberían apuntar a niveles de confianza altos que experimentan pocas personas. Por otro lado, hay que considerar que en esta escala la categoría ninguna vez fue poco plausible en casi todos los ítems. De hecho, los ítems 7 y 12 ubicados en esta escala mostraron parámetros b_1 muy bajos, menores a -1,9. Al revisar estos reactivos se puede concluir que esta respuesta indicaría que los sujetos tienen baja confianza de forma muy agresiva: "ninguna vez tuve confianza en mi propio desempeño" y "ninguna vez sentí que podía confiar en mí mismo". Por tanto, se recomienda cambiar la redacción de estos ítems para que la respuesta ninguna vez sea menos incómoda.

En la escala de Emocionalidad se deben construir ítems que presenten síntomas de ansiedad ante los exámenes en los que la opción ninguna vez no sea tan probable para personas con puntuaciones medias-bajas. De hecho, hubo reactivos en los que la respuesta ninguna vez fue muy probable hasta para población con niveles medios-altos de emocionalidad, por ejemplo: "temblé de nerviosismo", ya que el b_1 fue 0,73. El ítem en el que la categoría ninguna vez tuvo el límite de alta plausibilidad más bajo fue "me sentí nervioso/a"

con un b_1 de -0,64. Con base en lo anterior, se deberían agregar más ítems semejantes a este último (es decir, sintomatologías más comunes) y revisar el aporte del reactivo "temblé de nerviosismo".

En cuarto lugar, hay que mencionar las ventajas de los modelos TRI sobre el modelo de la TCT. El análisis de los errores estándar de medición evidenció que en los modelos de TRI no se utiliza el supuesto de que la precisión de la medición es igual en todo el continuo (Zanon et al., 2016). La liberación de este supuesto es una ventaja que ofrecen los modelos de TRI, ya que permite detectar los intervalos de puntuaciones que no son bien medidos por el instrumento, con lo cual se abren muchas oportunidades de mejora. Además, la estimación del error estándar de medición particular de cada nivel evita la formulación de conclusiones erróneas sobre la puntuación latente de las personas. Por ejemplo, un sujeto con un nivel alto de preocupación tendrá asociado un intervalo de confianza muy amplio, por lo cual no se puede descartar la hipótesis de que su nivel de preocupación sea medio; en cambio, bajo el modelo de la TCT si se podía rechazar dicha hipótesis (Embretson, 1996).

También hay que mencionar que el análisis de las CCC proporcionó información muy valiosa sobre los ítems que no es abordada en la TCT, como el rango en el que una categoría es la más probable. En la TCT se pueden estudiar proporciones de selección de categorías, según los niveles de las puntuaciones observadas, pero no se puede concluir claramente cuál es la más probable, porque no se usan modelos de probabilidad y porque las proporciones observadas de una categoría son inestables a lo largo de las puntuaciones observadas.

En conclusión, el trabajo desarrollado aportó más evidencias de que la GTAI es un instrumento apropiado para la medición de la AE, pero que puede ser mejorado para obtener precisiones más altas en la medición de ciertos rangos de las puntuaciones. Con base en lo anterior, un trabajo a futuro es construir un grupo de ítems con los resultados obtenidos en este estudio y analizar si efectivamente se alcanzan puntuaciones con precisiones aceptables en los puntos del continuo en que se esperarían mejoras.

Por último, una de las debilidades del estudio es que se contó con una muestra relativamente pequeña (184 personas), no obstante, el estudio de simulaciones indicó que el uso del MRG era pertinente, considerando las limitaciones en las estimaciones de las discriminaciones de las CCO y, además, los análisis estadísticos no mostraron problemas asociados a tamaños de muestra bajos, como parámetros con valores inesperados (Brown, 2006). En cuanto a las fortalezas del estudio se puede mencionar que se trabajó con una escala muy consolidada y que se analizó desde un enfoque que no se había considerado. A partir de lo anterior se encontraron debilidades del instrumento que habían sido ocultadas por el uso de los modelos tradiciones (Heredia et al., 2008; Piemontesi et al., 2012) y, se resaltó la importancia del uso de los modelos de TRI para el análisis de escalas.

Referencias

Auné, S. E., Abal, F. J., & Atorresi, H. F. (2020). Análisis psicométrico de una escala de ayuda con el modelo de respuesta graduada. *Psykhe*, *29*(2), 1–14. https://doi.org/10.7764/psykhe.29.2.1472

Baker, F. B., & Kim, S. (2004). *Item Response Theory: Parameter Estimation Techniques*. Springer.

Bonaccio, S., Reeve, C. L., & Winford, E. C. (2012). Text anxiety on cognitive ability test can result in differential predictive validity of academic performance. *Personality and Individual Differences, 52*, 497–502. https://doi.org/10.1016/j.paid.2011.11.015

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford press.

Cassady, J., & Johnson, R. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270–295. https://doi.org/10.1006/ceps.2001.1094

Cea, M. A. (2002). *Análisis multivariable: Teoría y práctica en la investigación social.* Síntesis.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical *Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chin, E., Williams, M. H., Taylor, J. E., & Harvey, S. H. (2017). The influence of negative affect on test anxiety and academic performance: An examination of the tripartite model of emotions. *Learning and Individual Differences, 54*, 1–8. https://doi.org/10.1016/j.lindif.2017.01.002

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Furlan, L. (2006). Ansiedad ante los exámenes. ¿Qué se evalúa y cómo? *Evaluar, 6*, 32–51. https://doi.org/10.35670/1667-4545.v6.n1.533

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hannon, B. (2016). General and non-general intelligence factors simultaneously influence SAT, SAT-V, and SAT-M performance. *Intelligence, 59*, 51–63. https://doi.org/10.1016/j.intell.2016.07.002

Heredia, D., Piemontesi, S., Furlan, L., & Hodapp, V. (2008). Adaptación del inventario alemán de ansiedad frente a los exámenes: GTAI-A. *Evaluar, 8*, 46–60. https://doi.org/10.35670/1667-4545.v8.n1.504

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponente. *Zeitschrift für Pädagogische Psychologie, 5*, 121–130.

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress, & Coping, 10*(3), 219–244. https://doi.org/10.1080/10615809708249302

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-X2item fit index for the graded response model. *Asia Pacific Education Review, 12*, 89–96. https://doi.org/10.1007/s12564-010-9082-4

Keith, N., Hodapp, V., Schermelleh-Engel, K., & Moosbrugger, H. (2003). Cross-sectional and longitudinal confirmatory factor models for the german test anxiety inventory: A construct validation. *Anxiety, Stress, & Coping, 16*(3), 251–270. https://doi.org/10.1080/1061580031000095416

Martínez, R., Hernández, M. J., & Hernández, M. V. (2014). *Psicometría*. Alianza Editorial.

Matteucci, M., & Stracqualursi, L. (2008). Student assessment via graded response model. *Statistica, 66*(4), 435–447. https://doi.org/10.6092/issn.1973-2201/1216

Maydeu-Olivares, A., & Joe, H. (2006). Limited- and full-Information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49,* 305–328. https://doi.org/10.1080/00273171.2014.911075

Mowbray, T., Jacobs, K., & Boyle, C. (2015). Validity of the german test anxiety inventory (TAI-G) in an Australian sample. *Australian Journal of Psychology, 67*, 121–129. https://doi.org/10.1111/ajpy.12058

Muñíz, J. (2000). *Teoría clásica de los tests*. Editorial Pirámide.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298. https://doi.org/10.1177/0146621603027004004

Owens, M., Stevenson, J., Hadwin, J. A., & Norgate, R. (2012). Anxiety and depression in academic performance: An exploration of the mediating factors of worry and working memory. *School Psychology International, 33*(4), 433–449. https://doi.org/10.1177/0143034311427433

Piemontesi, S., Heredia, D., & Furlan, L. (2012). Propiedades psicométricas de la versión en español revisada del Inventario Alemán de Ansiedad ante Exámenes (GTAI-AR) en universitarios argentinos. *Universitas Psychologica, 11*(1), 177–186. https://doi.org/10.11144/Javeriana.upsy11-1.ppve

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Recuperado a partir de https://www.R-project.org/

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.

Rojas, L. (2021). *Mecanismos subyacentes a la asociación de la relación de la ansiedad ante los exámenes con el rendimiento en pruebas* [Tesis de doctorado, Universidad de Costa Rica]. https://www.kerwa.ucr.ac.cr/handle/10669/83636?locale-attribute=en

Rojas-Torres, L., & Ordóñez, G. (2019). Proceso de construcción de pruebas educativas: El caso de la Prueba de Habilidades Cuantitativas. *Evaluar, 19*(2), 15–29. https://doi.org/10.35670/1667-4545.v19.n2.25080

Rosseel, Y. (2012). *lavaan: An R Package for Structural Equation Modeling*. http://www.jstatsoft.org/v48/i02/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(1), 1–97.

Samejima, F. (1997). Graded Response Model. En W. J. van der Linden & R. K. Hambleton, *Handbook of Modern Response Item Theory*. Springer-Verlag.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*, 929–938. https://psycnet.apa.org/doi/10.1037/0022-3514.46.4.929

Szafranski, D. D., Barrera, T. L., & Norton, P. J. (2012). Test anxiety inventory: 30 years later. *Anxiety, Stress & Coping, 25*(6), 667–977. https://doi.org/10.1080/10615806.2012.663490

Ul Hassan, M., & Miller, F. (2019). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics - Simulation and Computation, 48*, 1–19. https://doi.org/10.1080/03610918.2019.1705344

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92–104. https://doi.org/10.1037/h0031332

Zanon, C., Hutz, C. S., Hanwook, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(18). https://doi.org/10.1186/s41155-016-0040-x

Zeidner, M. (1998). *Test Anxiety: The State of Art*. Kluwer Academic Publishers.

Fecha de recepción: Julio de 2021
Fecha de aceptación: Agosto de 2022