Réplica a la "Crítica del análisis de la validez de constructo de la Escala de Detección de alumnos con Altas Capacidades para Padres (GRS 2); réplica a Tourón et al. (2023)"

Reply to "Criticism of the analysis of Construct Validity of the Gifted Rating Scales (GRS 2) Parent Form in Spain; a reply to Tourón et al. (2023)"

https://doi.org/10.4438/1988-592X-RE-2024-406-638

Marta Tourón
https://orcid.og/0000-0001-5430-4198
Universidad Internacional de La Rioja - UNIR

Enrique Navarro-Asencio
https://orcid.og/0000-0002-3052-146X
Universidad Complutense de Madrid

Javier Tourón
https://orcid.og/0000-0001-8217-1556
Universidad Internacional de La Rioja - UNIR

Resumen

El objetivo de este trabajo es responder a la revisión crítica realizada por Martínez, J. A. (2024) del estudio de la validez de constructo de la Escala GRS 2 para padres en España, Tourón et al. (2023). La crítica se centra principalmente en el modelo factorial propuesto, que utiliza una aproximación reflectiva del Análisis Factorial Confirmatorio (AFC) para la medida de la percepción de las altas capacidades. Mientras que en la revisión el autor propone una aproximación formativa. En esta respuesta se lleva a cabo una aclaración conceptual del término *giftedness* y también se realiza una réplica de las observaciones realizadas a la metodología de validación, poniendo la atención en las diferencias entre modelos reflectivos y formativos y las características del AFC utilizado. Además, se aportan

Recibido: 27/05/2024 Aceptado: 07/06/2024

evidencias que justifican la definición de un modelo de medida reflectivo a través de la estimación de un modelo bifactorial. Este modelo trata de explicar las respuestas a los ítems a partir de la consideración, al mismo tiempo, de un factor general y un conjunto de factores específicos. Los resultados muestran la importancia de un factor general, pero sin evidencias consistentes de una unidimensionalidad estricta. La presencia de una estructura multidimensional con cuatro factores específicos (capacidad cognitiva, capacidad creativa, habilidades sociales y control emocional) aporta una parte importante a la explicación de la varianza común del modelo.

Palabras clave: escala de detección, altas capacidades, validez de constructo, análisis factorial confirmatorio, modelos reflectivos y formativos, modelo bifactorial.

Abstract

The objective of this work is to respond to the critical review carried out by Martínez, J. A. (2024) of the study on the construct validity of the Gifted Rating Scales (GRS 2) parent form in Spain, Tourón et al. (2023). The critique mainly focuses on the proposed factorial model, which uses a reflective approach of Confirmatory Factor Analysis (CFA) for measuring the perception of high abilities, while the author in the review proposes a formative approach. This response provides a conceptual clarification of the term giftedness and also replicates the observations made on the validation methodology, focusing on the differences between reflective and formative models and the characteristics of the CFA used. Additionally, evidence is provided to justify the definition of a reflective measurement model through the estimation of a bifactorial model. This model seeks to explain the responses to the items by considering both a general factor and a set of specific factors simultaneously. The results show the importance of a general factor, but without consistent evidence of strict unidimensionality. The presence of a multidimensional structure with four specific factors (cognitive ability, creative ability, social skills, and emotional control) significantly contributes to the explanation of the common variance of the model.

Keywords: gifted rating scales, high ability, construct validity, confirmatory factor analysis, reflective and formative models, bifactor model.

Introducción

Agradecemos el minucioso estudio y análisis de Martínez, J. A. a nuestro trabajo de validación de constructo sobre la Escala de detección para Padres de las GRS 2 (*Gifted Rating Scales 2*) en España, en la que se

critican «varios de los procedimientos metodológicos y de análisis de resultados descritos por Tourón et al. (2023) son cuestionables. Tanto el modelo factorial propuesto, basado en una visión reflectiva de la medición de las altas capacidades, como los índices de ajuste empleados para validar la estructura dimensional, de carácter aproximado, limitan seriamente la interpretación de los resultados».

Vamos a replicar a las observaciones de Martínez, J. A. siguiendo la secuencia de su propia crítica y aportando nuestros argumentos respecto al trabajo realizado.

Este texto trata de dar respuesta a la crítica en diferentes aspectos relacionados con el modelo y la metodología utilizada en Tourón et al. (2023) para validar la estructura de dimensiones de la GRS 2 en su versión para familias. Recordemos que este instrumento es una escala de detección de comportamientos dotados basada en un modelo multidimensional de las altas capacidades, concretamente tres dimensiones: capacidades cognitivas, capacidades creativas y artísticas y habilidades socioemocionales (Pfeiffer y Jarosewich, 2003). Los ítems incluidos, un total de 20, reflejan comportamientos o características indicativas de alta capacidad y que pueden ser observadas por un padre o madre fuera de un entorno educativo. La validación no puede entenderse sin la consideración del uso de las puntuaciones que produce el test. Y, en este caso, la escala GRS 2 para padres, proporciona información que complementa la aportada por los docentes para identificar esos comportamientos. Antes de comenzar con la respuesta propiamente dicha, conviene proporcionar algo de contexto.

Sobre el constructo y la terminología

Hay una cuestión terminológica previa que conviene precisar. La llamada *giftedness* en inglés –que es el idioma en el que se ha escrito la literatura más relevante sobre el tema– significa dotación, de modo que el *gifted* es el dotado. El uso que se le ha dado a estos términos en castellano, en la mayor parte de las ocasiones, ha sido *superdotación* y *superdotado*. Términos no del todo correctos y no correspondientes con la literatura al respecto. Por ello, Tourón (2023) ha propuesto repetidamente atender a la dimensión sustantiva del término y utilizar las palabras *alta capacidad*. De modo que la capacidad es la dimensión sustantiva de la *giftedness*.

Muchos autores actuales (ver p. e. los planteamientos de Renzulli, Gagné o Subotnik et al. en Pfeiffer et al., 2018) entienden la *giftedness* como un constructo social, como una denominación que se refiere a un constructo multidimensional que tiene manifestaciones conductuales diversas en los planos cognitivo, afectivo, social, etc.

Tourón (2020) señalaba que «el problema conceptual que se da en este campo de estudio es mayúsculo y tiene consecuencias indeseables para la identificación y la intervención, lo que no es exclusivo de nuestro país. En cierta medida, ocurre en el ámbito anglosajón dentro de la comunidad educativa, no investigadora. Una primera problemática se deriva de la equiparación de términos como *gifted*, que quiere decir dotado, y hacerlo equivaler a *superdotado* (su correlato en ingles sería *supergifted*, que no existe), y este último ligarlo a la obtención de un valor de CI, generalmente 130, en alguno de los tests al uso. El problema central aquí es que la etiqueta se hace equivaler a un "estado del ser", o a un rasgo o cromosoma de oro (Cf. Renzulli y Reis, 2018), de manera que unas personas lo poseen y otras no. Se convierte de este modo el hecho de ser *gifted* en una variable dicotómica, cuyas categorías se establecen arbitrariamente a partir de un punto de corte (...). "El término 'dotado' [gifted en el original] significa que uno es excepcional en algo y preferimos usar la palabra "dotado" como adjetivo (por ejemplo, él o ella es un/a pianista, escritor/a, etc. dotado/a), en vez de como un sustantivo (ella es dotada). También preferimos hablar de comportamientos dotados (adjetivo) en lugar de usar los 'dotados' para representar un estado de ser (p.185)" (...).

» Todos los modelos actuales (...) enfatizan la importancia del desarrollo a lo largo de la vida de la persona, estableciendo la relevancia del impacto del entorno en dicho desarrollo. (...). Por ello, parece sensato entender que esta conceptualización nos ha de poner en primer plano la necesidad de identificar los potenciales diversos de las personas para ayudarlas a convertirlos en talentos o competencias desarrolladas, en rendimiento (Cf. Tourón, 2012). Más aún, como señala Pfeiffer (2017), el término *gifted* es un constructo social, una denominación que utilizamos para referirnos a un grupo heterogéneo de personas que se caracterizan por tener una alta capacidad, un alto rendimiento o potencial para rendir» (p.17-18).

Es necesario precisar que las escalas GRS 2 (en nuestro caso la de padres, aunque hay otras dos de profesores) proceden de una concepción previa de la *giftedness,* como realidad multidimensional que

Recibido: 27/05/2024 Aceptado: 07/06/2024

tiene manifestaciones externas en la conducta o comportamiento de las personas. Este conjunto de características observables, típicas de las personas con altas capacidades, se traducen en ítems que los padres (en este caso) han de valorar. De este modo, debería esperarse que los ítems diseñados para medir una determinada dimensión se relacionen entre sí de manera más intensa que si se han elaborado para medir dimensiones distintas. De esta forma, las dimensiones del constructo se consideran variables o dimensiones latentes, causantes de las respuestas a los ítems, que se consideran efectos.

Así, la escala es la dimensión operativa, medible, aunque sea de modo imperfecto, que permite estimar los rasgos latentes no medibles. Esto es un enfoque claramente reflectivo. La concepción de lo medido es anterior y base del instrumento de medida.

Finalmente, es preciso señalar que esta escala, como muchas otras en su género, no pretende medir las *altas capacidades*, como ocurre con un test de capacidad, por ejemplo, sino que pretende recabar información de otras fuentes (los padres) que valoran el grado en el que determinadas conductas o características están presentes –y en qué grado– en la persona valorada.

La estructura de la escala y su construcción es posterior a la conceptualización del constructo que pretende valorar, pero no es un instrumento que pretenda definir qué es la *giftedness*. (ver p. e. Pfeiffer y Jarosewich, 2003; Tourón et al., 2024).

En el trabajo de Tourón et al. (2023) los autores se centran en aportar evidencias psicométricas de la validez de constructo de la traducción española de la escala, pero no han participado en la definición teórica y operativa del constructo. También debe considerarse que la respuesta a los ítems que incluye la escala es producto de la percepción que tienen las familias sobre determinados comportamientos de sus hijos. Y que la respuesta a cada uno de esos ítems tiene un formato ordinal, utilizando una escala Likert de 6 puntos para identificar la frecuencia de cada uno de esos comportamientos.

Sobre la validación

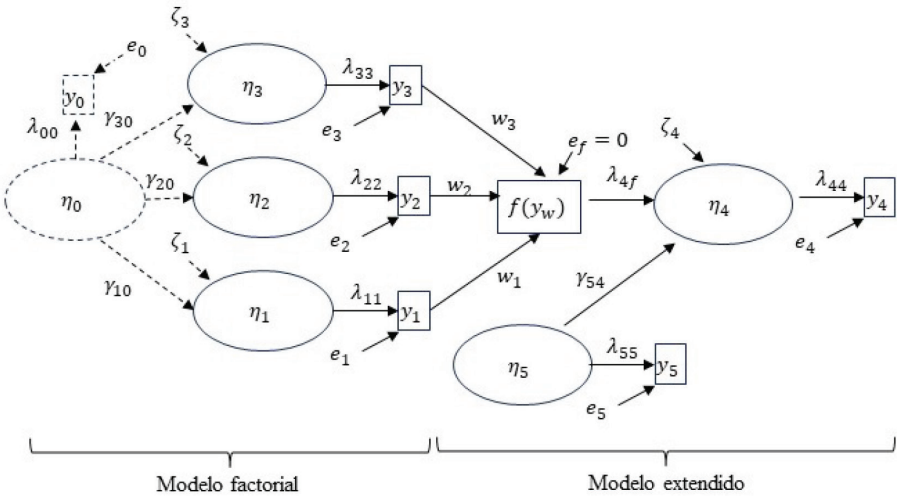
La validación de instrumentos tiene el propósito de recoger evidencias que apoyen la interpretación y uso de las puntuaciones de los test, como se señala en el manual de AERA, APA y NCME (2018). La validez se

refiere, como se sabe, al grado en el que la evidencia y la teoría apoyan las interpretaciones de las puntuaciones de los test asociadas a los usos previstos (p. 11). Las evidencias de validez pueden obtenerse de cinco fuentes: el contenido del test, el proceso de respuesta, la estructura interna, las relaciones con otras variables y las consecuencias de la aplicación del test. En el trabajo de Tourón et al. (2023) se aportan evidencias de la validez de la estructura interna de los ítems que forman la escala y su organización en dimensiones, pero tampoco es su propósito estimar las puntuaciones en los factores o baremar el instrumento.

La crítica metodológica se centra principalmente en la definición del modelo de medida utilizado para recoger evidencias de la relación entre los ítems y los constructos latentes. En el modelo confirmatorio utilizado en nuestro trabajo las respuestas a los ítems de la escala son considerados variables dependientes, es decir, son los efectos producidos por los factores latentes, que en nuestro modelo son las variables exógenas, las causas. Sin embargo, en la revisión se propone invertir el papel de ítems y factores, considerando los constructos formativos y no reflectivos.

En la revisión crítica, el autor argumenta que la aproximación más ajustada para definir la forma de calcular las puntuaciones factoriales es un modelo formativo, donde las puntuaciones en los diferentes ítems determinan el factor. De forma opuesta a la aproximación reflectiva utilizada en el trabajo, como se muestra en el modelo extendido de la siguiente figura:

FIGURA I. Modelo factorial y modelo extendido



Fuente: Martínez, J. A. (2024)

Recibido: 27/05/2024 Aceptado: 07/06/2024

La parte izquierda de la figura, el modelo factorial, representa nuestra estimación y el modelo de la derecha la propuesta de la crítica que se nos hace.

La utilización de una definición u otra es un asunto controvertido en la literatura psicométrica (Murray y Booth, 2018), optar por uno de los modelos implica asumir supuestos que van desde la definición teórica y operativa del constructo, hasta su validación empírica. Emplear una aproximación formativa conlleva tratar las dimensiones latentes como variables exógenas, es decir, como un efecto producto de la combinación de diferentes indicadores. Y, por tanto, no considerarlas un factor endógeno, la causa de que los sujetos respondan de una determinada manera a los diferentes ítems.

Murray y Booth (2018) mencionan 5 aspectos para diferenciar entre indicadores como causas o efectos. El primero, es la dirección de causalidad; en los modelos formativos, los ítems son la causa del factor latente y en los reflectivos son efectos. El segundo, son los cambios que los indicadores pueden producir en el constructo; en el modelo formativo cualquier cambio en el indicador producirá un efecto en el valor del constructo, por el contrario, en los reflectivos los cambios en un indicador no deben altear el nivel en el constructo. El tercer argumento se refiere a si el constructo es la causa común del conjunto de indicadores, cuestión que se produce en el caso del modelo reflectivo, donde los indicadores necesitan estar correlacionados. El cuarto, es si los indicadores tienen las mismas causas y efectos, un requisito que es necesario en los modelos reflectivos, pero no en los formativos. Finalmente, también mencionan la intercambiabilidad de los indicadores que, en principio, debe cumplirse en los modelos reflectivos.

Siguiendo a Zumbo (2006), la distinción reflectiva y formativa diferencia entre medidas e índices. En el primer caso, un cambio en la variable latente se refleja en cambios en los indicadores (las respuestas a los ítems que se consideran efectos). Y, en el segundo, el índice es una variable latente donde los cambios en los indicadores provocan cambios en el factor, son, por tanto, las causas. Este autor señala que realizar un ACP no es evidencia suficiente para contar con un índice (constructo formativo) y tampoco los procedimientos de AFC determinan que se cuente con una medida (constructo reflectivo). Sin embargo, aplicar un ACP a un constructo reflectivo no es lo apropiado. En el trabajo de Bollen y Lennox (1991) proponen que esta diferenciación debe explorarse

durante la validación de contenido, preguntando a los expertos por el carácter formativo o reflectivo de la medida. Incluso con procedimientos de pensamiento en voz alta durante una aplicación piloto.

¿Por qué un modelo reflectivo?

El modelo utilizado en Tourón et al. (2023) busca aportar evidencias psicométricas de la estructura de dimensiones de la escala y también de la calidad de los indicadores utilizados para su medida. En este contexto, el trabajo estudia distintos modelos para contrastar empíricamente diferente número de dimensiones y propone un modelo final de cuatro dimensiones, aunque también se prueba y muestra un buen ajuste un modelo con dos factores de segundo orden. Desde esta perspectiva, se asume la existencia de constructos no observables que determinan la forma de responder a los ítems que incluye un test. Y también que en esa respuesta hay parte explicada por el constructo latente y otra parte que es error de medida.

Debe tenerse en cuenta que el trabajo de Tourón et al. (2023) valida la estructura de dimensiones de la traducción del instrumento GRS 2, pero no han participado en el análisis inicial de la validez del contenido, que tiene el propósito de validar el conjunto de indicadores y asegurar que son una muestra de conductas de la red nomológica que define el constructo. Este tipo de validez suele realizarse a partir de un juicio de expertos. Aunque se asume que los ítems de la escala GRS 2 cumplen con la validez de contenido. y qué el propósito de los indicadores es representar las conductas del constructo de medida, pero que pueden hacer referencia a diferentes niveles del mismo factor.

La descripción hecha en la revisión crítica debe completarse con las siguientes puntualizaciones.

En primer lugar, Martínez, J. A. menciona que "en línea con la teoría clásica de los tests: la variación en la variable latente η_1 se manifiesta en una variación en el indicador observable, escalada con el parámetro λ_11, a lo que hay que añadir un error aleatorio e_1." (p. 5) De la siguiente forma:

$$y_1 = \lambda_{11}\eta_1 + e_1 \quad (1)$$

Recibido: 27/05/2024 Aceptado: 07/06/2024

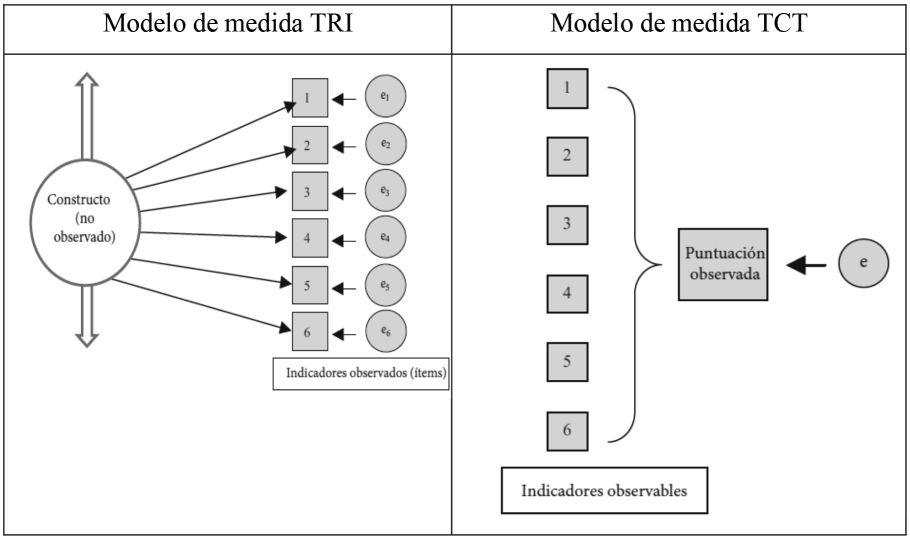
Sin embargo, esta aproximación para definir las respuestas a los ítems se ajusta más a los supuestos de la TRI que de la TCT. En la TCT o Teoría de la puntuación verdadera, el modelo pone la atención en la definición de la puntuación total en el test como resultado de la suma de la puntuación verdadera y un error de medida:

$$X=V+E \qquad (2)$$

La dispersión de esos errores es un indicador de la precisión de las puntuaciones observadas y puede estimarse mediante la fiabilidad. Las inferencias se hacen utilizando la puntuación total en el test (escala), que se calcula a partir de la información observada. En cambio, en la TRI el modelo define de forma explícita el comportamiento del constructo a medir, del rasgo latente. Son las respuestas a los ítems las que están determinadas, en términos de probabilidad, por el nivel que tienen un sujeto en el constructo.

El modelo de la TCT no incluye información sobre la relación entre el constructo latente y la forma de responder a los ítems. Solo el concepto de puntuación verdadera hace esa referencia, es parte no observada. La diferencia entre los dos modelos de medida puede verse en el siguiente gráfico:

FIGURA II. Modelos de medida en la TCT y la TRI



Fuente: Adaptado de Wu, Tam y Jen (2016).

En el modelo TRI las flechas señalan el efecto del constructo en la probabilidad de responder de una forma u otra al ítem. Es, por tanto, el constructo el que determina todo el patrón de respuestas (modelo reflectivo). Los errores (e) también se representan con círculos, son variables que no se observan directamente, y señalan la influencia en la respuesta de otros factores desconocidos, distintos al constructo que queremos medir. El error es un término asociado a cada ítem y determina la capacidad que tiene el constructo para explicar la respuesta al mismo.

En la TCT, tanto las respuestas a los ítems como la puntuación total del test son datos observados. El total del test se calcula como una agregación de las respuestas a los ítems (suma de aciertos, promedios, etc.) (modelo formativo) y el error está asociado a esta puntuación total y no a cada uno de los ítems. La diferencia entre estos dos modelos refleja la aproximación reflectiva y formativa utilizada para definir el constructo y también las diferencias entre un Análisis Factorial Confirmatorio (AFC) y un Análisis de Componentes Principales (ACP).

En segundo lugar, Martínez, J. A. presenta una aproximación lineal del análisis factorial, y la perspectiva utilizada en el trabajo de Tourón et al. (2023), debido al carácter ordinal de los ítems y la falta de normalidad multivariada, es no lineal. Así, el modelo para definir la respuesta a los ítems estima la puntuación en el constructo necesaria para situarse entre las diferentes opciones de respuesta. Se empleó la matriz de correlaciones policóricas como elemento informativo en el análisis factorial. Al utilizar este tipo de correlación se asume la existencia de una variable continua subyacente (Jöreskog, 1994) y las respuestas politómicas observadas se consideran manifestaciones de los encuestados que superan un determinado número de puntos de corte o umbrales dentro de ese continuo. En este sentido, el modelo estima esos umbrales (*thresholds* "τ") y define las respuestas observadas en las diferentes categorías ordinales a través de variables continuas latentes. Concretamente, para un ítem i con un número de categorías c=0,1,2…, C la variable latente y* se define de forma que:

$$yi = c \text{ if } \tau c < yi* < \tau c+1 \qquad (3)$$

donde τc, τc+1 son los umbrales que determinan los puntos de corte en la variable continua latente subyacente, que suelen estar espaciados en intervalos de diferente amplitud. Considerando este supuesto, la correlación de interés para el modelo está entre estas variables continuas

Recibido: 27/05/2024 Aceptado: 07/06/2024

(correlación policórica). El procedimiento de análisis suele realizarse en tres pasos (Jöreskog,1990; Muthén, 1984). En las primeras dos etapas se estiman los umbrales y las correlaciones policóricas y, en la tercera etapa, estos valores se ajustan en un modelo hipotetizado utilizando algún método de estimación, principalmente DWLS. Los parámetros del modelo se obtienen minimizando la función que compara la información estimada con los datos del modelo.

En tercer lugar, en la revisión Martínez, J. A. menciona una variable latente de orden superior, a modo de factor general, «conceptualización habitual en esta área de conocimiento es la primera, tal y como se muestra, por ejemplo, en Pfeiffer et al. (2008), donde se plantea una concepción multidimensional de las altas capacidades con un factor subyacente "g" o factor de capacidad general. Hay que reconocer que Tourón et al. (2023) son muy cautos y no afirman claramente este hecho, pero subrepticiamente parecen hacerlo cuando calculan una "media total de la escala", como posteriormente explicaremos» (p. 6). Aquí es conveniente mencionar que en la validación no se prueba en ningún momento un factor general, sino que se presenta una estructura con dos factores de segundo orden. El primero determina las puntuaciones en el factor cognitivo y el factor creativo y, el segundo, para el factor social y el factor de control emocional. Además, la media del conjunto de ítems calculada en la sección de estadísticos descriptivos no pretende mostrar la existencia de ese factor general, es solo un resumen de los datos. También conviene recordar que el objetivo del trabajo consistió en aportar evidencias de la validez de constructo, pero no estimar las puntuaciones en las dimensiones.

En cuarto lugar, relacionado con lo anterior, Martínez, J. A. menciona que «proponer un modelo multidimensional (…) que refleja un factor g subyacente, implica desde el punto de vista de la medida, que el factor g puede medirse con el "mejor" indicador de la "mejor" dimensión, es decir, con un único ítem(…) Esto es compatible, como no podría ser de otro modo, con la visión reflectiva sobre la medición, donde los ítems de una variable latente se pueden considerar intercambiables, lo que indica que quitando un ítem no se altera el significado de la variable latente» (p. 7).

En este argumento se hace referencia a la intercambiabilidad de los indicadores utilizados para definir el constructo y también para la definición de los factores de segundo orden. No queda claro si Martínez, J. A. alude aquí al supuesto de paralelismo de las puntuaciones de

la TCT (Lord y Novick, 1968) o el paralelismo conceptual de los indicadores que definen el constructo, que asume que intercambiar uno por otro no altera el significado del constructo (Borsboom et al., 2004). En el primer tipo de paralelismo, para estimar la fiabilidad de un test con datos empíricos, los ítems se consideran pequeñas partes paralelas que reflejan la puntuación verdadera y se utilizan para la evaluación de la consistencia interna. Lograr medidas completamente paralelas implica que las medias, las desviaciones típicas y errores de medida son equivalentes, algo muy complejo de lograr en la práctica. Si no se logra la misma dispersión nos situamos ante medidas tau-equivalentes. Otro caso de paralelismo, las medidas esencialmente tau-equivalentes, permite también que las medias varíen entre las diferentes partes añadiendo una constante. Finalmente, las medidas congenéricas son las menos restrictivas y permiten también que esas medias difieran añadiendo o multiplicando por una constante. En nuestra validación, no puede asumirse una estructura completamente paralela porque los ítems de un mismo constructo son una muestra de los comportamientos, pero pueden reflejar diferentes intensidades de esa medida. En cualquier caso, los modelos de AFC permiten contar con medidas que no cumplen con el paralelismo estricto, es decir, con pesos factoriales distintos. Además, la base de este análisis es la consistencia interna de los indicadores que definen el constructo, donde se espera que haya correlación. Sin embargo, la aproximación formativa en un ACP no necesita que los indicadores que determinan el factor estén correlacionados, además se asume que no son intercambiables. El AFC permiten contar con indicadores que tienen una correlación heterogénea con los factores latentes. Aquí, la fiabilidad compuesta (ω) se calcula a partir de las cargas factoriales para producir estimaciones más precisas que los procedimientos de consistencia interna, como el α de Cronbach.

En los factores de segundo orden sí puede ser más ajustado definir un modelo formativo si los diferentes factores asociados a la alta capacidad se consideran partes de un factor general. Considerando los resultados de la validación, los dos factores de segundo orden (Capacidades Cognitivo-Creativas y Habilidades Socio-Emocionales) definidos no pueden considerarse intercambiables. Como muestran los resultados de correlación (0,53), comparten el 25% de variabilidad. No obstante, como ya hemos mencionado, el factor general no se probó en nuestra validación.

Recibido: 27/05/2024 Aceptado: 07/06/2024

En quinto lugar, Martínez, J. A. realiza una analogía entre capacidades intelectuales y físicas. Aquí es conveniente apuntar que la escala GRS es un instrumento de percepción, no mide directamente la capacidad intelectual, sino que se infiere a partir de la observación de determinados comportamientos. Además, los indicadores para medir capacidades físicas son observables y sin error, a no ser que el instrumento utilizado para recoger la medida no funcione correctamente. En cambio, la percepción estará determinada en parte por el constructo, pero también puede estar influida por algún factor no medido (error).

Es posible que para la medida de la capacidad física sea necesaria una única medida de cada indicador y una combinación de todos ellos para estimar la dimensión. Sin embargo, para poder medir con precisión dimensiones latentes a partir de indicadores basados en la respuesta a ítems, es necesaria más de una medida de la misma conducta. Por ejemplo, para medir la capacidad aritmética en competencia matemática de Educación Primaria, puede incluir un único ejercicio para sumar dos cantidades o incluir varios con el mismo propósito en un mismo examen. Más ítems aumentarán la fiabilidad de la medida.

En sexto lugar, Martínez, J. A. también propone reducir el número de indicadores para probar esta cuestión, sin embargo, como ya hemos mencionado, los indicadores pueden reflejar diferentes niveles del constructo al no considerarse medidas paralelas y también dependerá de la complejidad conceptual del constructo medido. Insistimos en que los indicadores pueden considerarse intercambiables, en el modo que son conductas determinadas por el mismo constructo, pero para poder medir con mayor precisión son necesarios varios indicadores. En caso contrario, las puntuaciones tendrían mucho error de medida.

Martínez, J. A. también menciona que «Tourón et al. (2023) parecen considerar como algo positivo cuando aluden a las correlaciones altas entre cada ítem y el resto señalando la *"homogeneidad del conjunto de datos"*. Es más, hay ejemplos en la propia literatura sobre altas capacidades y escalas GRS donde se vislumbra un efecto halo que afecta obviamente a la validez (ej. Jabůrek et al., 2020).» (p. 12). No obstante, asumiendo la posible causa común de las respuestas a los indicadores y la consistencia interna de los mismos, esta homogeneidad es una característica de los modelos reflectivos. La respuesta a los indicadores está provocada, en parte, por el factor latente, pero el modelo asume error de medida. Ese error muestra que hay otros factores que pueden

determinar esa respuesta. Martínez, J. A. también menciona que «lo que probablemente está sucediendo es que esa batería de indicadores está midiendo variables latentes diferentes, es decir, no son la manifestación de una única variable latente, sino de varias.» (p. 12). En este sentido, considerando los valores de r^2 de los ítems y los promedios de varianza explicada en cada factor (AVE), la parte explicada por las dimensiones latentes es mayor que el posible efecto de otros factores no tenidos en cuenta. No obstante, aportar evidencias de variables que puedan determinar el sesgo podría ser otra evidencia más de la validez de constructo.

Finalmente, en séptimo lugar, Martínez, J. A. realiza una crítica al procedimiento seguido en la metodología, incluyendo la utilización de un Análisis Factorial Exploratorio, como etapa previa al AFC y los índices de ajuste utilizados. Se citan los trabajos de Hayduk (2014a y 2014b) como justificación de la crítica, mencionado que «el análisis factorial exploratorio es incapaz de detectar la estructura real de los datos, es decir, de identificar el modelo que ha generado esos datos empíricos» (p. 11), argumentando que χ^2 puede utilizarse incluso con tamaños de muestra grandes y que si no pasa este test no pueden interpretarse el resto de los parámetros. Es conveniente mencionar aquí que el AFE, realizado como primera etapa, no tiene un propósito confirmatorio, se emplea para extraer información inicial sobre el número de dimensiones y recoger evidencias de la consistencia de la estructura de dimensiones con el modelo confirmatorio. En la validación de la escala GRS 2, se prueban diferentes modelos confirmatorios, de tres y cuatro dimensiones y se toman decisiones considerando los índices de modificación, por ejemplo, el cambio del ítem 17 del factor cognitivo al factor creativo.

La afirmación sobre el índice de ajuste χ^2 es demasiado contundente a nuestro parecer. Los procedimientos de validación de constructo en el ámbito de la medición educativa y la psicometría apuntan la necesidad de interpretar con precaución este índice de ajuste en el contexto de los AFC. El estadístico χ^2 pone a prueba la hipótesis nula de que la matriz de covarianzas observadas en la población es equivalente a la producida por el modelo, sin embargo, en Ciencias Sociales, cualquier modelo se considera una aproximación a la realidad y, por tanto, esa hipótesis nula con un ajuste exacto es inviable (Bentler y Bonett, 1980; Schermelleh-Engel et al., 2003). Aunque Hayduk (2014b) indica que los valores de χ^2 no están afectados por el tamaño muestral cuando el modelo está

correctamente especificado, pero considerando esa particularidad de los modelos utilizados en Ciencias Sociales no sería posible. El mismo autor apunta que aumentar la muestra incrementa el poder de χ^2 para detectar problemas en la especificación del modelo. La cuestión aquí es saber si, cuando el valor de χ^2 resulta significativo, es realmente un indicador suficiente para descartar el modelo y poder cometer un error Tipo I o si ese modelo es la mejor aproximación a la realidad y el posible sesgo asumible. Wang y Wang (2020) argumentan que la probabilidad de rechazar un modelo incrementa sustancialmente cuando el tamaño de las muestras aumenta, incluso cuando las diferencias entre las matrices de varianzas-covarianzas observadas y estimadas son pequeñas. Señalan que el valor de χ^2 aumenta al no cumplir con el supuesto de normalidad multivariada y las distribuciones de respuesta a los ítems son asimétricas o están afectadas por algún tipo de curtosis. Y también que el índice χ^2, cuando el número de variables en el modelo se incremente. Razones suficientes para no utilizarlo de forma exclusiva.

En nuestro caso, la estimación utiliza una ponderación de la matriz de varianzas y covarianzas para los valores de χ^2 y los errores típicos. Recordemos que, al no cumplir con el supuesto de normalidad multivariada, se utilizó WLSMV y la matriz de correlaciones policórica. Este método de estimación es una versión de DWLS (Muthén, 1984), pero aplicando la corrección robusta de media y varianza a los mínimos cuadrados ponderados y un cambio en la escala (son los denominados WLSMV o ULSMV). Y, aunque en el trabajo se presentan los valores de χ^2, con este tipo de datos no conviene utilizarlo (Finney y DiStefano, 2013). En el trabajo de Shi et al. (2018) se estudió cómo afecta el tipo de estimación, el tamaño de la muestra o la complejidad del modelo a diferentes índices de χ^2 y muestran un mejor funcionamiento general de los índices robustos.

Bentler y Bonett (1980) propusieron el uso de índices de ajuste incremental, como CFI y TLI, para calcular la cantidad de información que se gana cuando se comparan modelos. Estos índices evalúan el grado en el que un modelo estimado es mejor que un modelo nulo (modelo en el que todas las variables observadas no están correlacionadas) en su capacidad para reproducir la matriz de varianzas-covarianzas observadas. Y, sobre todo, las medidas de ajuste absoluto, donde se comprueba si el modelo definido se corresponde con los datos empíricos. Los índices de ajuste absoluto, como RMSEA o SRMR, no utilizan un modelo nulo

de referencia, pero realizan una comparación implícita con un modelo saturado que reproduce de forma exacta la matriz de varianzas-covarianzas de las variables observadas (Hu y Bentler, 1999). El índice RMSEA señala la falta de ajuste entre el modelo especificado en la población y el SRMR estima la raíz cuadrada del promedio de los residuos.

Considerando lo anterior, todas las aportaciones que permitan comprobar de forma empírica la validez del constructo de la escala GRS 2 refuerzan la calidad psicométrica del instrumento. Con el propósito de afianzar las evidencias a favor del modelo propuesto originalmente en el trabajo de Tourón et al. (2023), se ha llevado a cabo un modelo bifactorial (Holzinger y Swineford, 1937; Chen et al., 2006) para comprobar la existencia de un factor general que pueda ser una causa común de las respuestas a los ítems, junto con los cuatro factores definidos.

Modelo bifactorial

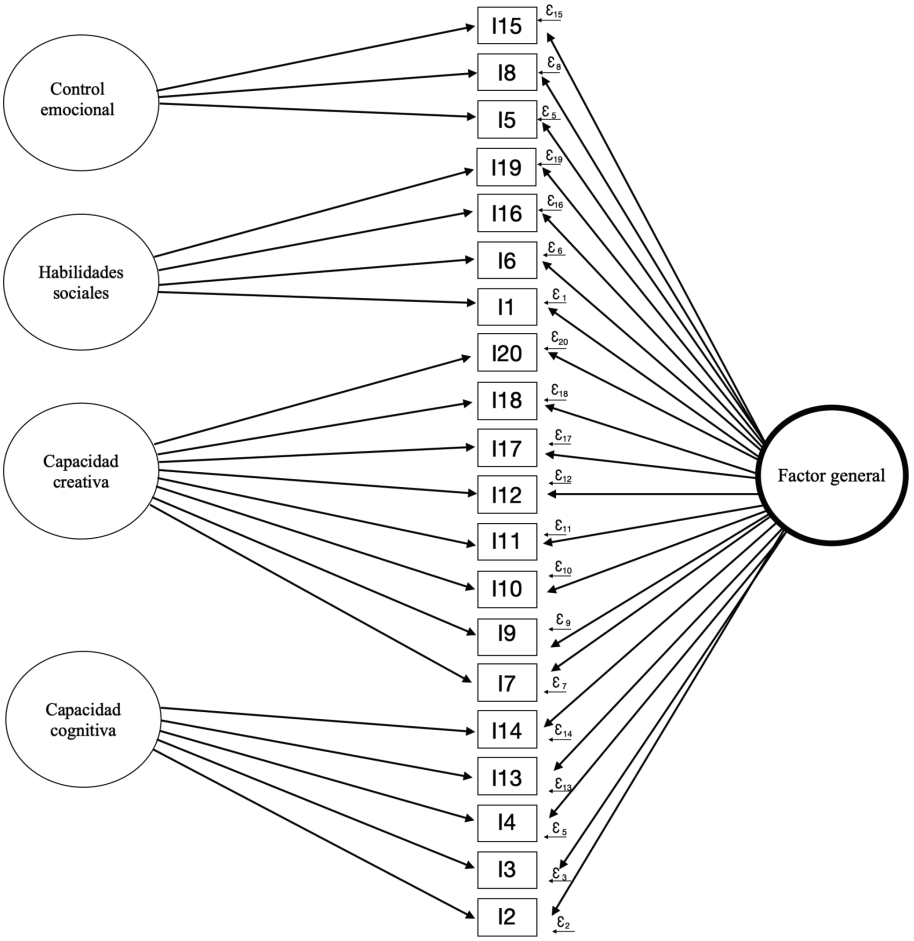
Esta propuesta define que la varianza compartida entre las respuestas a los ítems en dos partes, por un lado, la explicada por un factor general y aquella que está determinada por un grupo de factores específicos, que pueden ser del mismo dominio. Por tanto, plantea la hipótesis de un factor general para explicar la varianza común y, al mismo tiempo, también la existencia de múltiples factores que tienen un impacto independiente en la explicación de esa varianza (ver Figura III). Dado que los factores específicos se interpretan como la varianza contabilizada por encima y más allá del factor general, se asume que las relaciones entre los factores generales y específicos son ortogonales (no correlacionadas).

Chen y Zhang (2018) señalan que la capacidad de estudiar los factores específicos independientemente del factor general es importante para comprender mejor las afirmaciones teóricas. Por ejemplo, si un factor específico propuesto no explicara una cantidad sustancial de varianza más allá del factor general, se observarían cargas factoriales pequeñas y no significativas en el factor específico, así como una varianza no significativa del factor específico en el modelo bifactorial. Cuestión que evidenciaría que el factor específico no proporciona una explicación a la varianza más allá del factor general.

Para llevar a cabo una evaluación del modelo, después de obtener el ajuste del modelo tanto para modelos unidimensionales como

Recibido: 27/05/2024 Aceptado: 07/06/2024

FIGURA III. Modelo bifactorial



Fuente: Elaboración propia.

bifactoriales, se pueden comparar directamente mediante la diferencia entre los índices CFI (ΔCFI), ya que el modelo unidimensional está jerárquicamente anidado dentro del modelo bifactorial (Reise, 2012). Este índice se calcula como:

$$\Delta CFI = CFI\_M1 - CFI\_M0 \quad (4)$$

Donde CFI_M1 es igual al valor CFI obtenido para el modelo 1, y CFI_M0 es igual al valor CFI obtenido para el modelo 0. Este índice es más estable en condiciones diferentes, como el tamaño de la muestra, la cantidad de error, el número de factores y el número de ítems y se recomiendan valores iguales o inferiores a .01 para confirmar la equivalencia (Meade et al., 2008).

Para la fiabilidad se pueden emplear cuatro versiones del coeficiente omega (o fiabilidad compuesta): omega total para el factor general (ωt), omega para cada subdimensión (ωs) y los coeficientes omega jerárquicos, que también pueden calcularse para el factor general (ωht) y para cada subdimensión (ωhs). El coeficiente omega McDonald (1999) es un tipo de fiabilidad basada en los resultados del análisis factorial, adecuado cuando se cuenta con medidas congenéricas (distintos pesos factoriales), que estima la proporción de la varianza observada atribuible a los factores del modelo:

$$\omega t = \frac{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Cog}\right)^2 + \left(\sum \lambda_{Crea}\right)^2 + \left(\sum \lambda_{Soc}\right)^2 + \left(\sum \lambda_{Emo}\right)^2}{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Cog}\right)^2 + \left(\sum \lambda_{Crea}\right)^2 + \left(\sum \lambda_{Soc}\right)^2 + \left(\sum \lambda_{Emo}\right)^2 + \left(\sum_{i=1}^{n} e_i\right)} \quad (5)$$

En el denominador están todas las fuentes de varianza del modelo, la varianza común producida por todos los factores del modelo, general y subdimensiones, más la varianza específica de los errores. El numerador incluye solo esas las fuentes de varianza común. También puede calcularse para cada subdimensión:

$$\omega s = \frac{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Subdimensión}\right)^2}{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Subdimensión}\right)^2 + \left(\sum_{i=1}^{n} e_i\right)} \quad (6)$$

En este caso, los pesos factoriales y los errores son de los ítems correspondientes a la subdimensión. Y para determinar la proporción de la varianza total de las puntuaciones, debida únicamente al factor general, se utiliza el coeficiente omega jerárquico, que se calcula dividendo la suma de las cargas factoriales del factor general al cuadrado por la varianza total, considerando varianza común y error:

Recibido: 27/05/2024 Aceptado: 07/06/2024

$$\omega h = \frac{\left(\sum \lambda_G\right)^2}{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Cog}\right)^2 + \left(\sum \lambda_{Crea}\right)^2 + \left(\sum \lambda_{Soc}\right)^2 + \left(\sum \lambda_{Emo}\right)^2 + \left(\sum_{i=1}^{n} e_i\right)} \quad (7)$$

La varianza debida a las subdimensiones se considera aquí parte del error de medida, un coeficiente de 0,8 más señala que las puntuaciones pueden considerarse esencialmente unidimensionales, considerando el factor general como la principal fuente de varianza. Además, podemos saber cuánto aporta cada una de las dimensiones a la varianza común, una vez controlado el factor general (Reise, 2012).

$$\omega hs = \frac{\left(\sum \lambda_{Subdimensión}\right)^2}{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Subdimensión}\right)^2 + \left(\sum_{i=1}^{n} e_i\right)} \quad (8)$$

Como en el caso de la varianza total, solo se emplean los pesos factoriales y errores de los ítems que componen la subdimensión. Como referencia pueden emplearse los puntos de corte mencionados por Smits et al. (2014), donde los valores iguales o superiores a 0,3 pueden considerarse importantes, los inferiores hasta 0,2 moderados y por debajo de 0,2 bajos.

Desde esta perspectiva, el índice de varianza común explicada (ECV) se utiliza para contrastar la unidimensionalidad de las escalas a partir de las cargas factoriales del factor general y de las subdimensiones (Reise et al., 2013), de la siguiente forma:

$$ECV = \frac{\left(\sum \lambda_G\right)^2}{\left(\sum \lambda_G\right)^2 + \left(\sum \lambda_{Cog}\right)^2 + \left(\sum \lambda_{Crea}\right)^2 + \left(\sum \lambda_{Soc}\right)^2 + \left(\sum \lambda_{Emo}\right)^2} \quad (9)$$

Es, por tanto, la proporción de varianza explicada por el factor general dividida por la varianza explicada por ese factor y las subdimensiones. Valores altos señalan una gran importancia del factor general, pero establecer puntos de corte no es sencillo, pero se sugieren valores de 0,7 o superiores para considerar la unidimensionalidad (Rodríguez et al., 2016). Este indicador también se puede calcular en el nivel del ítem para identificar aquéllos donde la influencia del factor general es muy potente, autores como Stucky y Edelen (2014) proponen valores superiores a 0,8 o 0,85.

Recibido: 27/05/2024 Aceptado: 07/06/2024

Para interpretar el índice de unidimensionalidad ECV, se recomienda calcular también el porcentaje de correlaciones no contaminadas (PUC) (Rodríguez et al., 2016). Este índice, junto al ECV, informan sobre el posible sesgo al forzar datos multidimensionales en un modelo unidimensional calcula cuántas correlaciones modera los efectos que puede tener el factor general. PUC puede definirse como el número de correlaciones no contaminadas dividido por el número de correlaciones únicas:

$$\mathrm{PUC} = \frac{\frac{I_G * (I_G - 1)}{2} - \left[\frac{I_{s1} * (I_{s1} - 1)}{2} + \frac{I_{s2} * (I_{s2} - 1)}{2} + \cdots + \frac{I_{sn} * (I_{sn} - 1)}{2}\right]}{\frac{I_G * (I_G - 1)}{2}} \qquad (10)$$

I_G es el número de ítems que cargan en el factor general, I_s es el número de ítems que cargan en cada factor específico. Cuando los valores de PUC están por encima de 0,8 los valores de ECV no son muy relevantes porque señala que las cargas factoriales del modelo unidimensional se aproximarán a las obtenidas en el factor general del modelo bifactorial. En cambio, si los valores de PUC están por debajo de 0,8, deben obtenerse valores de ECV superiores a 0,6 para considerar la unidimensionalidad. Y si los valores son muy altos (>,90), pueden obtenerse estimaciones unidimensionales insesgadas incluso cuando se obtiene un valor de ECV bajo (Reise, 2012).

Otros indicadores que puede ayudar a evidenciar la calidad del modelo es el índice H de replicabilidad del constructo. Este índice permite valorar si el conjunto de ítems que representa cada variable latente es adecuado y se calcula de la siguiente forma:

$$\mathrm{H} = \frac{1}{\left[1 + \frac{1}{\sum_{i=1}^{n} \frac{\lambda^2}{1 - \lambda^2}}\right]} \qquad (11)$$

Es, por tanto, un sumatorio de la tasa de la varianza explicada por la variable latente, dividida por el error de medida, es decir, la proporción de la variabilidad del constructo que se explica por sus propios indicadores. Los valores de 0,7 o más señalan que la variable latente

Recibido: 27/05/2024 Aceptado: 07/06/2024

está bien definida por sus indicadores y tendrá estabilidad en diferentes estudios (Rodríguez et al., 2016).

Para finalizar, se analizan las cargas factoriales del modelo bifactorial y se comparan los resultados del factor general con los obtenidos en el modelo puramente unidimensional (M3, ver Tabla I). A partir de las diferencias entre los valores de las cargas de cada ítem de los dos modelos se calcula el sesgo relativo de los parámetros (SRP):

$$SRP = \frac{\lambda_G - \lambda_{UNIDIM}}{\lambda_G} * 100 \quad (12)$$

El promedio de este índice informa sobre el sesgo que puede cometerse al ajustar un modelo unidimensional cuando no se cumple ese supuesto. Los valores superiores al 15% señalarían ese posible sesgo (Rodríguez et al., 2016). Además, como señalan Ferrando y Lorenzo-Seva (2018), si en el modelo bifactorial el factor general acumula cargas altas y en las dimensiones, en cambio, el promedio de sus cargas no supera el 0,3, puede ser otro indicador de la estructura unidimensional de la medida.

Resultados

En primer lugar, se incluyen los valores de ajuste del modelo bifactorial comparados con algunos de los probados en el trabajo de Tourón et al. (2023). Concretamente, se presentan los resultados del modelo original (M1), el modelo unidimensional (M3), el modelo propuesto (M6) y su versión con dos factores de segundo orden (M8). Por último, el nuevo modelo bifactorial (M9).

El modelo bifactorial, considerando las diferencias de los índices de ajuste incremental y absoluto con el modelo propuesto (M6), pueden considerarse equivalentes en su capacidad explicativa. También muestra un mejor ajuste que el modelo puramente unidimensional (M3).

En segundo lugar, los valores de omega absoluto y jerárquico para el factor general y las subdimensiones se presentan en la siguiente tabla:

Los valores de omega (ω) señalan la fiabilidad de cada factor y el coeficiente jerárquico (ωh) muestra la parte atribuible a cada uno en la explicación de la varianza total. Como se observa, el valor de ωh del factor general está por debajo de 0,7 pero evidencia una aportación

TABLA I. Índices de ajuste de los modelos y diferencia entre el modelo bifactorial el propuesto en Tourón et. al. (2023)

Índices	**M1**	**M3**	**M6**	**M8**	**M9**	**Diferencia (Δ) M9-M8**
AFC	3	1				
χ^2	2048	6833	1596	1601	1760	
gl	167	170	164	165	150	
p	<,001	<,001	<,001	<,001	<,001	
χ^2/gl	12,263	40,194	9,732	9,703	11,733	
SRMR	0,086	0,154	0,074	0,074	0,081	0,007
RMSEA	0,101	0,189	0,089	0,089	0,098	0,009
CFI	0,968	0,867	0,976	0,976	0,973	-0,003
TLI	0,964	0,851	0,972	0,972	0,966	-0,006
GFI	0,978	0,91	0,983	0,983	0,981	-0,002

Fuente: Elaboración propia

TABLA II. Omega y Omega jerárquico de los factores (general y específicos)

	F. General	**Cognitivo**	**Creativo**	**Social**	**Emocional**
ω	0,941	0,880	0,921	0,826	0,789
ωh	0,669	0,427	0,566	0,495	0,509

Fuente: Elaboración propia.

relevante. También los factores aportan la explicación de la varianza total, con valores de ωh alrededor del 50%.

Comprando los resultados ω y ωh del factor general se observa si el 66,9% de la variabilidad está determinada por ese factor, el otro 27,2% está provocado por las diferencias de los factores específicos. El resto, un 5,9% se atribuye, por tanto, al error de medida.

En tercer lugar, el índice de unidimensionalidad ECV es de 0,71, es decir, un 71% de la varianza común está explicada por el factor general y el otro 29% por las subdimensiones. Recordemos que valores de 0,8 señalan la unidimensionalidad y cercanos a 0,7 podrían ser indicadores

Recibido: 27/05/2024 Aceptado: 07/06/2024

también de esa posibilidad. Sin embargo, el índice ECV para los ítems solo detecta uno con un valor por encima de 0,8, es el ítem 17, como muestra la siguiente tabla III.

Además, el porcentaje de correlaciones sin contaminar (PUC) es de 0,75. Valor inferior a 0,8, que, en combinación con el ECV de 0,669, pone en duda la presencia única de un factor general.

TABLA III. Índice de varianza común explicada el ítem (I_ECV)

Ítem	I_ECV	Ítem	I_ECV
I1	0,469	I11	0,419
I2	0,455	I12	0,632
I3	0,727	I13	0,471
I4	0,433	I14	0,498
I5	0,172	I15	0,256
I6	0,507	I16	0,469
I7	0,172	I17	0,805
I8	0,780	I18	0,485
I9	0,209	I19	0,193
I10	0,395	I20	0,270

Fuente: Elaboración propia.

En cuarto lugar, el índice H de replicabilidad del constructo señala la capacidad de representación que tienen el conjunto de ítems para definir cada factor. Los resultados se muestran en la siguiente tabla:

Se logra, en todos los casos, valores cercanos o superiores a 0,7. Por tanto, el 70% o más de la variabilidad de cada factor latente está determinado de indicadores que lo componen y puede considerarse aceptable.

TABLA IV. Índice de replicabilidad del constructo (H)

	F. General	Cognitivo	Creativo	Social	Emocional
H	0,869	0,697	0,865	0,678	0,721

Fuente: Elaboración propia.

Finalmente, como muestra la Tabla V, los pesos factoriales muestran cargas moderadas para el factor general, entre 0,3 y 0,6, pero todas significativas. Los valores en las subdimensiones, aunque similares, son algo más altos y también resultan significativas, como muestra la siguiente tabla. En promedio, tanto el factor general como las subdimensiones, tienen cargas factoriales considerables, superando el punto de corte propuesto de 0,3. En el caso del factor general se encuentra muy cerca de

TABLA V. Pesos factoriales estandarizados, R^2 e índice de sesgo relativo de los parámetros (SRP) del modelo bifactorial (M9)

	λ*					λ*		
Ítem	**F. General**	**Cognitiva**	**Creativa**	**Social**	**Emocional**	**R^2**	**Unidim.**	**SRP**
I2	0,602	0,659				0,797	0,696	0,034
I3	0,632	0,387				0,549	0,602	0,156
I4	0,504	0,577				0,587	0,574	0,047
I13	0,595	0,631				0,751	0,692	0,139
I14	0,404	0,406				0,328	0,399	0,066
I7	0,339		0,745			0,670	0,740	0,054
I9	0,407		0,792			0,793	0,829	1,183
I10	0,577		0,714			0,842	0,897	0,169
I11	0,570		0,671			0,775	0,854	1,037
I12	0,438		0,334			0,303	0,508	0,555
I17	0,461		0,227			0,264	0,467	0,498
I18	0,520		0,536			0,558	0,701	0,160
I20	0,432		0,711			0,692	0,773	0,163
I1	0,446			0,475		0,424	0,431	0,012
I6	0,517			0,510		0,527	0,489	0,010
I16	0,548			0,583		0,639	0,527	0,038
I19	0,341			0,698		0,604	0,390	0,013
I5	0,362				0,793	0,761	0,386	0,348
I8	0,544				0,289	0,379	0,452	0,144
I15	0,390				0,665	0,595	0,394	0,789
Promedio	0,481	0,532	0,591	0,567	0,582	0,592	0,590	0,281

*Todos los parámetros son significativos (p<,001).

Recibido: 27/05/2024 Aceptado: 07/06/2024

0,5 y en las subdimensiones algo por encima de ese valor y con niveles muy similares entre ellas.

Observando los valores de los pesos factoriales de la tabla anterior, en 5 ítems de los 20 que componen la escala, el valor es mayor en el factor general que en las subdimensiones. Son el ítem 3, 6, 8, 12 y 17, pertenecientes a las diferentes subdimensiones del modelo. Y el modelo bifactorial (M9) logra explicar aproximadamente el 60% de variabilidad de los datos.

Respecto al sesgo relativo de los parámetros (SRP), el valor promedio está cerca del 30%, por tanto, al exceder el límite del 15%, queda constancia de las diferencias sustanciales en los efectos de ese factor general en los dos modelos.

Conclusiones

Los modelos reflectivos no pueden considerarse superiores a los formativos, o viceversa. Ambas pueden ser alternativas en el estudio de la validez de constructo (Bollen y Diamantopoulos, 2017). No obstante, en el caso concreto de la escala GRS la opción reflectiva utilizada en el estudio de la validez de constructo de Tourón et al. (2023), se ajusta a la definición teórica y operativa del mismo. Considerando los resultados del modelo bifactorial, se confirma la presencia de un factor general como causante de gran parte de la variabilidad de las respuestas a los ítems y también de una estructura multidimensional de factores específicos que aporta otra parte a la varianza común del modelo (capacidad cognitiva, capacidad creativa, habilidades sociales y control emocional).

El valor del índice de unidimensionalidad (ECV) y el porcentaje de correlaciones sin contaminar (PUC), con valores inferiores a 0,8 en ambos casos, no muestra la presencia clara de un único de un factor general para explicar las respuestas a los ítems. No obstante, estos resultados muestran la importancia de ese factor, pero también la influencia de las subdimensiones en la explicación de las diferencias. Y, como señalan los coeficientes omega jerárquicos de las subdimensiones, todos cercanos o superiores a 0,5, pueden considerarse efectos importantes, Smits et al. (2014). Raise et al. (2013) apuntan que con valores de PUC inferiores a 0,7, un valor de ECV del factor general superiores a 0,6 y un omega jerárquico superior a 0,7 sugieren la presencia de cierta

multidimensionalidad, aunque no descartan totalmente la interpretación de la escala como unidimensional.

Las diferencias de intensidad existentes entre los pesos factoriales de cada factor demuestran el carácter congenérico de las medidas, pero los resultados señalan la presencia de una causa común que determina gran parte de la variabilidad de la matriz de varianzas-covarianza. Ese factor general identificado en el modelo factorial determina aproximadamente el 70% (ECV=0,71) de la variabilidad de la varianza común y, por tanto, la intercambiabilidad de los indicadores, considerados conductas provocadas por el mismo constructo, puede mantenerse. Esta cuestión es clave para definir un modelo de medida como reflectivo (Murray y Booth, 2018). El resto de la varianza común, un 30% aproximadamente, está producida por las diferencias en los resultados de las subdimensiones.

Los pesos factoriales han resultados significativos en todos los casos, tanto en el factor general como en las subdimensiones. Y los índices ECV de cada ítem (ver Tabla III) han mostrado que el factor general determina una parte de las respuestas a los ítems, pero solo en el caso del ítem 17 esa aportación supera el 80%. En los ítems 3 y 8 el factor general explica aproximadamente el 75%. En 10 de los ítems, la aportación se sitúa entre el 41%-60% y en los siete ítems restantes es aproximadamente del 20%-40%. Por tanto, en la mayoría de los ítems, la influencia del factor general se combina con el impacto de las subdimensiones.

Además, los índices H de replicabilidad del constructo muestran buenos resultados, con valores cercanos o superiores a 0,7. Por tanto, el conjunto de indicadores que componen cada variable latente explica una variabilidad suficiente y tendrá estabilidad en diferentes estudios (Rodríguez et al., 2016).

Considerando los valores de las cargas factoriales, en el modelo bifactorial muestran un tamaño suficiente tanto en el factor general como en las subdimensiones, con valores promedio cercanos o superiores a 0,5, mostrando la importancia del modelo completo (Ferrando y Lorenzo-Seva, 2018). También el índice de sesgo relativo de los parámetros (SRP) apunta que las cargas factoriales del factor general en el modelo bifactorial son diferentes a las del modelo unidimensional. Por tanto, no considerar la estructura multidimensional produce sesgo en las estimaciones.

Agradecemos toda la revisión crítica llevada a cabo y esperamos haber respondido a la mayor parte de las objeciones planteadas de manera satisfactoria. Tendremos en cuenta las nuevas vías de consideración que se abren como consecuencia de esta discusión y reanálisis de nuestros

Recibido: 27/05/2024 Aceptado: 07/06/2024

datos. Finalmente agradecemos al editor de la revista que haya aceptado incluir estos trabajos a raíz del nuestro original, pues entendemos que estos debates y diferencias de posturas metodológicas hacen progresar y mejorar el trabajo científico.

Referencias bibliográficas

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas (M. Lieve, Trans.). AERA*. American Educational Research Association.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Bollen, K.A. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110(2)*, 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A Comparison of Bifactor and Second-Order Models of Quality of Life. *Multivariate Behavioral Research*, *41*(2), 189–225. https://doi.org/10.1207/s15327906mbr4102_5

Chen, F.F., & Zhang, Z., (2018). Bifactor Models in Psychometric Test Development. En P. Irwing, T. Booth y D.J. Hughes, *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*, 325-345. https://doi.org/10.1002/9781118489772.ch12

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and psychological measurement*, 78(5), 762-780. https://doi.org/10.1177/0013164417719308

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller

(Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). IAP Information Age Publishing.

Hayduk, L. A. (2014a). Seeing perfectly-fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measuremen*t, *74*(6), 905-926. https://doi.org/10.1177/0013164414527449

Hayduk, L. A. (2014b). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC: Medical Research Methodology, 14,* 124 https://doi.org/10.1186/1471-2288-14-124

Holzinger, K.J., and Swineford, F. (1937). The Bi-factor method. *Psychometrika 2,* 41–54. https://doi.org/10.1007/BF02287965

Hu, L. y Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6(1),* 1-55. https://doi.org/10.1080/10705519909540118

Jabůrek, M., Ťápal, A., Portešová, Š., Pfeiffer, S. I. (2020). Validity and Reliability of Gifted Rating Scales-School Form in Sample of Teachers and Parents – A Czech Contribution. *Journal of Psychoeducational Assessment*, *39*(3), 361–371. https://doi.org/10.1177/0734282920970718

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*, 387–404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment (1st ed.).* Psychology Press. https://doi.org/10.4324/9781410601087

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568-592. https://doi.org/10.1037/0021-9010.93.3.568

Murray, A.L. & Booth, T. (2018). Causal Indicators in Psychometrics. InP. Irwing, T. Booth and D.J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing*. https://doi.org/10.1002/9781118489772.ch7

Muthén, B. (1984). A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Pfeiffer, S. (2017). *Identificación y evaluación del alumnado con altas capacidades: Una guía práctica*. La Rioja: UNIR Editorial.

Pfeiffer, S. I., & Jarosewich, T. (2003). *GRS: Gifted Rating Scales*. Psychological Corporation.

Pfeiffer, S. I., Petscher, Y., & Kumtepe, A. (2008). The Gifted Rating Scales-School Form: A Validation Study Based on Age, Gender, and Race. *Roeper review, 30*(2), 140–146. https://doi.org/10.1080/02783190801955418

Pfeiffer, S. I., Shaunessy-Dedrick, E., & Foley-Nicpon, M. (Eds.). (2018). *APA handbook of giftedness and talent. American Psychological Association*. https://doi.org/10.1037/0000038-000

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. http://dx.doi.org/10.1080/00273171.2012.715555

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*, *73*(1), 5-26. https://doi.org/10.1177/0013164412449831

Renzulli, J. S., & Reis, S. M. (2018). The three-ring conception of giftedness: A developmental approach for promoting creative productivity in young people. In S. I. Pfeiffer, E. Shaunessy-Dedrick, & M. Foley-Nicpon (Eds.), *APA Handbook of Giftedness and Talent* (pp. 185–199). American Psychological Association. https://doi.org/10.1037/0000038-012

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online, 8*(2), 23-74.

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining Chi-Square Test Statistics Under Conditions of Large Model Size and Ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 924-945. https://doi.org/10.1080/10705511.2018.1449653

Smits, I. A., Timmerman, M. E., Barelds, D. P., & Meijer, R. R. (2014). The Dutch symptom checklist-90-revised. *European Journal of Psychological Assessment*, *31*(4), 263-271. https://doi.org/10.1027/1015-5759/a000233

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. En S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 183-206). Routledge/Taylor & Francis.

Tourón, J. (2012). *¿Superdotación o alta capacidad?* Recuperado de https://www.javiertouron.es/superdotacion-o-alta-capacidad/

Tourón, J. (2020). Las altas capacidades en el sistema educativo español: reflexiones sobre el concepto y la identificación: Concept and Identification Issues. *Revista de Investigación Educativa*, *38*(1), 15–32. https://doi.org/10.6018/rie.396781

Tourón, J. (2023). ¿Puede una escuela inclusiva ignorar a sus estudiantes con altas capacidades? *enTERA2.0, 10*, 24-46.

Tourón, M., Navarro-Asensio, E., Tourón, J. (2023). Validez de Constructo de la Escala de Detección de alumnos con Altas Capacidades para Padres, (GRS 2), en España. *Revista de Educación*, *402*, 55-83. https://doi.org/10.4438/1988-592X-RE-2023-402-595

Tourón, M., Tourón, J., & Navarro-Asencio, E. (2024). Validación española de la Escala de Detección de altas capacidades, «Gifted Rating Scales 2 (GRS 2-S) School Form», para profesores. *Estudios Sobre Educación*, *46*, 33-55. https://doi.org/10.15581/004.46.002

Wang, J. y Wang, X. (2020). *Structural Equation Modeling: Applications Using Mplus, Second Edition*. John Wiley & Sons Ltd.

Wu, M., Tam, H. P., & Jen, T. H. (2016). *Educational measurement for applied researchers. Theory into Practice*. https://doi.org/10.1007/978-981-10-3302-5

Zumbo, B. D. (2006). 3 validity: foundational issues and statistical methodology. En C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics, 45-79*. Elsevier Science. https://doi.org/10.1016/S0169-7161(06)26003-6

Información de contacto: Marta Tourón. Universidad Internacional de La Rioja - UNIR. Avd. de La Paz, 137. 26006 Logroño La Rioja. E-mail: marta.tporto@unir.net

Recibido: 27/05/2024 Aceptado: 07/06/2024