Contribución al conocimiento en Investigación Educativa a partir de estudios de validación de escalas: una reflexión crítica sobre el trabajo de M. Tourón et al. (2023) y el análisis metodológico de Martínez-García (2024)

Contribution to knowledge in Educational Research through scale validation studies: a critical review on the work of M. Tourón et al. (2023) and the methodological analysis by Martínez-García (2024)

https://doi.org/10.4438/1988-592X-RE-2024-406-637

Fernando Martínez Abad
https://orcid.org/0000-0002-1783-8198
Universidad de Salamanca

José Carlos Sánchez Prieto
https://orcid.org/0000-0002-8917-9814
Universidad de Salamanca

Resumen

Existe en la literatura científica un gran volumen de investigaciones que proponen el diseño y validación de escalas de medida en el ámbito de las ciencias de la educación. Este trabajo presenta, desde el punto de vista de la investigación educativa aplicada, un análisis crítico de la validación de la escala de detección de alumnos con altas capacidades (GRS 2) para padres publicada por M. Tourón et al. (2023) en Revista de Educación, teniendo en cuenta la crítica metodológica realizada por Martínez-García (2024) a dicha validación. La presente propuesta pretende sumarse a este diálogo mediante el análisis de ambas publicaciones y la formulación de cuestiones de mejora, centrándose

Recibido: 01/06/2024 Aceptado: 01/06/2024

especialmente en la fundamentación teórica de los modelos, la formulación de compuestos reflectivos y formativos, y la adecuación del uso de indicadores de bondad de ajuste alternativos al Chi Cuadrado en el modelo de análisis factorial confirmatorio. Se observa la aplicación de ciertos procedimientos muy asentados en los estudios de validación psicométrica dentro de la investigación educativa que, además de no promover la comprensión de las bases teóricas en las que se fundamenta la escala, dificultan su adecuado empleo posterior por parte de investigadores aplicados en estudios diagnósticos o experimentales. Así, los argumentos analizados ponen de relieve la necesidad de replantear, a nuestro juicio, algunas de las prácticas habituales en este tipo de estudios que disminuyen su relevancia y las implicaciones de los resultados obtenidos para el desarrollo de la teoría y la práctica educativa.

Palabras clave: validación de escalas, análisis factorial exploratorio, análisis factorial confirmatorio, modelos de ecuaciones estructurales, modelos formativos.

Abstract

In the scientific literature, there is a considerable volume of research proposing the design and validation of measurement scales in the field of educational sciences. This work presents, from the perspective of applied educational research, a critical analysis of the validation of the detection scale for high-ability students (GRS 2) for parents published by M. Tourón et al. (2023) in Revista de Educación, taking into account the methodological critique made by Martínez-García (2024) of said validation. The present proposal aims to contribute to this dialogue by analyzing both publications and formulating improvement issues, focusing especially on the theoretical foundation of the models, the formulation of reflective and formative compounds, and the adequacy of using alternative goodness-of-fit indicators to Chi Square in the confirmatory factor analysis model. Certain procedures deeply established in psychometric validation studies within educational research are observed, which, besides not promoting the understanding of the theoretical foundations underlying the scale, hinder its subsequent appropriate use by applied researchers in diagnostic or experimental studies. Thus, the analyzed arguments highlight the need to reconsider, in our view, some of the common practices in this type of studies that diminish their relevance and the implications of the results obtained for the development of educational theory and practice.

Keywords: scales validation, explanatory factor analysis, confirmatory factor analysis, structural equations modelling, formative models.

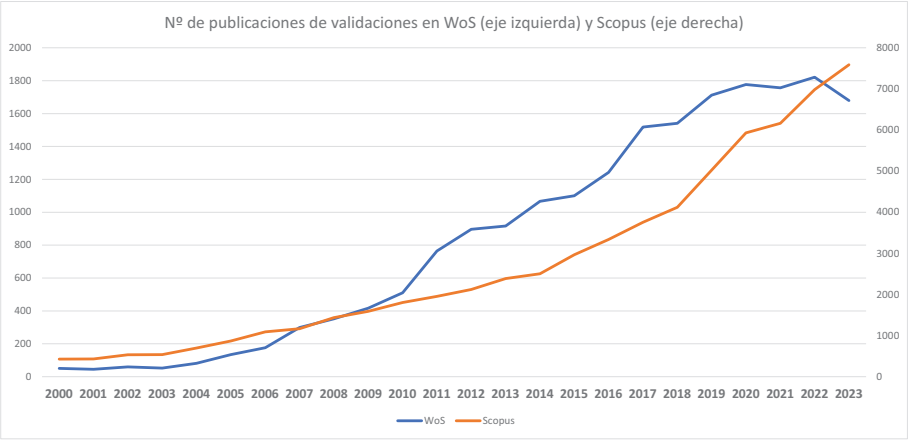
Introducción

Los estudios psicométricos orientados al diseño y validación de escalas e instrumentos conforman una línea de investigación muy popular y profusa en Investigación Educativa. Esto se refleja en una amplia presencia en publicaciones científicas del área de Ciencias Sociales y de la Educación, como ilustra la Figura I. De hecho, el aumento del interés por el diseño y validación de escalas de medida es exponencial, con un importante crecimiento en el número de publicaciones científicas desde 2005 en ambas bases de datos.

Como evidencia de esta amplia presencia, cada vez son más las revistas que argumentan esta casuística como justificación en el rechazo de propuestas, llegando a estar recogido en la normativa para los autores en algunas de ellas (e.g., Educación XX1, 2024).

Un buen ejemplo de este tipo de trabajos es el estudio de M. Tourón et al. (2023), publicado recientemente en la Revista de Educación. El estudio aborda una validación psicométrica en la población española de la Escala de Detección de alumnos con Altas Capacidades para Padres

FIGURA I. Evolución del número de publicaciones de validaciones de escalas en WoS y Scopus



Nota: Búsqueda en WoS en fuentes del Área ***Education & Educational Research***, y en Scopus en fuentes del Área ***Social Sciences***. Cadena de búsqueda: *(validation OR design OR psychometric) AND (scale OR instrument)*.

(GRS 2), incluyendo una exploración inicial mediante un análisis factorial exploratorio (AFE) y, a partir de estos resultados previos, la validación de ocho modelos dimensionales diferentes propuestos mediante un análisis factorial confirmatorio (AFC), hasta la selección del modelo más adecuado a los datos.

El proceso de diseño y validación estadística de escalas es potencialmente valioso para la comunidad científica, ya que posibilita la medida adecuada de rasgos o constructos a una población objetivo en estudios de carácter aplicado, ya sean exploratorios, correlacionales o experimentales. Así, bajo nuestro punto de vista, la presentación de los resultados de estudios psicométricos debe documentar clara y sistemáticamente toda la información que permita a otros investigadores educativos aprovechar y utilizar su potencial.

En este contexto, y teniendo en cuenta que el análisis crítico de Martínez-García (2024) se centra fundamentalmente en las decisiones metodológicas tomadas por M. Tourón et al. (2023) en el proceso de validación estadística, la presente revisión pretende contribuir al debate poniendo el foco principal en la aportación práctica que supone para los investigadores educativos aplicados tanto la validación de la escala GRS 2 propuesta por M. Tourón et al. (2023) como las críticas de Martínez-García (2024).

Este documento se divide en cuatro apartados principales. En primer lugar, se presenta una reflexión en torno a la aportación que realiza el trabajo de M. Tourón et al. (2023), junto con las críticas de Martínez-García (2024), al desarrollo científico en el ámbito de las Ciencias de la Investigación. Posteriormente, se analiza la pertinencia de las decisiones metodológicas tomadas en el artículo de validación psicométrica, poniendo el foco en si estas decisiones contribuyen a una mejor comprensión de la escala validada, fomentando el empleo de la misma en estudios aplicados y, finalmente, un avance en el conocimiento sobre el estudio de las altas capacidades. El tercer bloque de este trabajo se centra en desgranar y discutir las principales críticas de Martínez-García (2024), de nuevo bajo un punto de vista más práctico-aplicado que a partir del enfoque teórico-básico que desarrolla el autor del análisis crítico. Finalmente, se incluye un bloque a modo de conclusión en el que se reflexiona en torno a lo generalizado en la investigación educativa de las dudosas prácticas aquí identificadas, y la necesidad de una toma de conciencia sobre las mismas que lleve al desarrollo de trabajos con una orientación que no se centre principalmente en la publicabilidad.

Recibido: 01/06/2024 Aceptado: 01/06/2024

Contribución al corpus científico en investigación educativa

Cabe destacar inicialmente que valoramos muy positivamente los procedimientos técnicos y decisiones estadísticas implementadas en el trabajo de M. Tourón et al. (2023). Tal y como destaca Martínez-García (2024), el estudio de validación realiza un esfuerzo importante por desarrollar un procedimiento sistemático y conforme a los estándares y usos más habituales en la investigación psicométrica. No obstante, encontramos algún obstáculo importante que dificulta la trazabilidad y aplicabilidad de los resultados obtenidos y presentados.

En primer lugar, a pesar de que la escala GRS 2 está publicada originalmente en lengua inglesa, y que los ítems fueron traducidos para llevar a cabo el proceso de validación, llama la atención que M. Tourón et al. (2023) no presentan el listado completo de los ítems traducidos. Esta omisión dificulta la adecuada comprensión y seguimiento de las decisiones tomadas en relación a los modelos dimensionales probados en el trabajo, a la vez que imposibilita el empleo de la escala en estudios aplicados. Dado que la escala original pertenece a una editorial (MHS), lo más probable es que no haya sido posible publicar la escala completa. Esta cuestión la indican los autores de manera indirecta cuando se refieren a la traducción de los ítems. Cuando menos, este tipo de prácticas no son las más adecuadas para contribuir al avance del conocimiento científico, la ciencia abierta y el desarrollo de futuros trabajos académicos.

Por otro lado, a pesar del esfuerzo realizado por M. Tourón et al. (2023) en definir todos los modelos dimensionales plausibles a partir del conjunto de ítems y probar su ajuste, los autores no dedican ese mismo esfuerzo en justificar la pertinencia y sentido teórico de cada uno de los modelos validados.

Si bien es cierto que los autores incluyen en las conclusiones del trabajo una reflexión teórica en torno al modelo finalmente seleccionado, la omisión de una justificación más amplia sobre todos los modelos probados no sólo dificulta al lector una mejor comprensión de la escala GRS 2 y de su estructura dimensional, sino que limita sensiblemente el alcance del estudio al reducir el análisis de la relación entre teoría y praxis.

A este respecto, habría sido de utilidad para la comunidad científica una mayor discusión sobre las implicaciones de los resultados para la conceptualización de la inteligencia y la relación entre sus distintos componentes, resultando especialmente llamativa la división de las tres

Recibido: 01/06/2024 Aceptado: 01/06/2024

dimensiones iniciales en cuatro durante el AFE a las que se le añadieron dos de segundo orden durante el AFC.

Como añadido, no está de más recordar en este punto los riesgos asociados a la definición de la estructura dimensional de una escala teniendo en cuenta exclusivamente los datos empíricos (Brown, 2015; MacCallum et al., 1992; Schmitt et al., 2018). En este sentido, consideramos que puede existir un sesgo por sobreajuste en los resultados del AFC.

En suma, el trabajo de M. Tourón et al. (2023) no incluye el contenido de los ítems de la escala GRS 2 ni una justificación clara de los 8 modelos dimensionales validados en el AFC aplicado. Estas dos carencias dificultan la labor hermenéutica del artículo e impiden su reproducibilidad al quedar incompleto el proceso de modelado de las dimensiones teóricas de la escala. Con un buen desarrollo teórico, el artículo podría haber ido más allá de la presentación de una herramienta valida y fiable para la detección de altas capacidades, aportando información de interés para la comprensión de los componentes de la inteligencia.

En conjunto, estas limitaciones suponen una barrera para que otros investigadores y profesionales del ámbito educativo puedan seguir avanzando en la detección y estudio de alumnado con altas capacidades.

Procedimiento y decisiones metodológicas

Tal y como hemos indicado anteriormente, tras realizar una exploración inicial de la estructura dimensional mediante AFE, M. Tourón et al. (2023) proponen y comprueban el ajuste de 8 modelos dimensionales diferentes en el AFC. Aunque se sigue el procedimiento habitual en este tipo de análisis, en la redacción del trabajo no queda clara la conexión entre los resultados del AFE y los modelos propuestos en el AFC. En este punto estamos de acuerdo parcialmente con la crítica de Martínez-García (2024) sobre el uso del AFE en el trabajo, al no quedar claro qué aporta el AFE aplicado a las decisiones tomadas en la aplicación del AFC. En este sentido, consideramos que la aplicación del AFE en este artículo es contingente: existe una teoría previa sólida sobre la estructura factorial de la escala, y las evidencias obtenidas en el AFE no conducen a la toma de decisiones sobre qué modelos testear en el AFC.

Aunque luego abordaremos esta cuestión en mayor profundidad, debemos decir que no estamos de acuerdo con las críticas de Martínez-

Recibido: 01/06/2024 Aceptado: 01/06/2024

García (2024) sobre el empleo de índices de ajuste complementarios al contraste Chi Cuadrado. Al igual que los estadísticos de tamaño del efecto son tremendamente valiosos en la adecuada interpretación de los contrastes de hipótesis bivariados (Cohen, 1988; Grissom & Kim, 2011), los indicadores de ajuste absoluto e incremental complementarios a Chi Cuadrado permiten realizar un análisis más fino y claro del grado de similitud entre el modelo empírico y los modelos nulo y saturado.

Como se ha mencionado anteriormente, una de las limitaciones principales de este tipo de estudios es el carácter puramente empirista de la toma de decisiones. Al respecto de los modelos probados en el AFC, los autores incluyen modelos con errores correlacionados tomando en cuenta la información aportada por los índices de modificación. Esta práctica no es la más recomendable, ya que puede dar lugar a modelos artificialmente sobreajustados y difícilmente interpretables en la práctica (MacCallum et al., 1992; Schmitt et al., 2018). Teniendo en cuenta que la escala original está previamente validada en una población similar, otra decisión que genera dudas es el cambio de dimensión que sufre el ítem 17 en alguno de los modelos. Dado que los autores no informan sobre los pesos factoriales de los modelos AFE aplicados, y no se conoce la redacción de los ítems, no se aportan evidencias suficientes que justifiquen esta decisión.

Una crítica fundamental de Martínez-García (2024) es la relativa a la decisión no justificada de M. Tourón et al. (2023) de considerar las dimensiones de la escala GRS 2 como reflectivas. Los autores no plantean la posibilidad de que los ítems, o el constructo de segundo orden probado en alguno de los modelos, puedan conformar una estructura formativa. Debemos tener en cuenta que la presencia de estructuras dimensionales formativas en los estudios psicométricos del ámbito de las ciencias de la educación es residual. En este sentido, consideramos importante dar a conocer en el campo de la investigación educativa las características y propiedades de los factores formativos, en relación a la naturaleza y supuestos de los factores reflectivos.

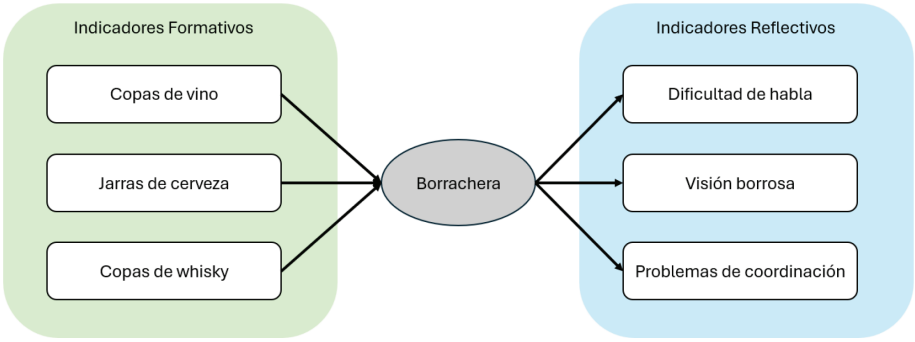
Un factor reflectivo es aquel en el que las variables latentes se consideran compuestos en los que los indicadores son el resultado de dicha variable, de esta manera, un cambio en el compuesto reflectivo provoca un cambio en todos los indicadores. Por otro lado, en el caso de los factores formativos las variables latentes son consideradas compuestos en los que los indicadores son su causa de forma que un cambio en el

compuesto no va necesariamente acompañado de un cambio en todos los indicadores (Starsted et al., 2016). Esta diferencia en la relación entre compuesto e indicador se representa gráficamente mediante la dirección de las flechas que unen el compuesto con sus indicadores, mientras que en los compuestos reflectivos las flechas van del compuesto hacia sus indicadores (al igual que sucede en los modelos CB-SEM), en el caso de los compuestos formativos las flechas van en la dirección contraria.

Un ejemplo clásico de la diferencia entre compuestos reflectivos y formativos es el *ejemplo del borracho* (Figura II) (Chin, 1998). En este caso, si modelásemos la borrachera de manera reflectiva este tendría indicadores como "visión doble", "dificultad en el habla" o "problemas de coordinación" de manera que un incremento en la borrachera generaría un aumento en sus indicadores. De manera opuesta, si la borrachera fuese un compuesto formativo esta incluiría indicadores como "número de cervezas", "número de copas de vino" o "número de copas de whisky", en este caso un aumento de la borrachera no implica un aumento en todos los indicadores.

La cuestión del uso de modelos que incluyen compuestos formativos constituye un punto abierto de debate, que se entremezcla con la disputa en la comunidad científica sobre la validez del análisis confirmatorio de compuestos (ACC) que sólo puede llevarse a cabo mediante la aplicación de modelos de ecuaciones estructurales basados en mínimos cuadrados parciales (PLS-SEM), una técnica de naturaleza predictivo-causal (Hair et al., 2011). Esta cuestión tiene implicaciones que se abordarán más adelante.

FIGURA II. Diferencia entre indicadores reflectivos y formativos



Fuente: Elaboración propia.

En lo que se refiere a la crítica realizada por Martínez-García (2024), resulta difícil precisar si la escala validada por los autores incluye compuestos reflectivos o formativos, dado que no se aborda esta cuestión en el artículo ni, como ya se ha mencionado, se incluye el listado completo de los ítems que la componen para que el lector pueda juzgar esta cuestión por sí mismo. No obstante, en nuestra opinión, a juzgar por alguno de los ítems mencionados y teniendo en cuenta las estructuras teóricas planteadas, existe la posibilidad de que se trate de compuestos formativos al menos en alguno de los casos. En resumen, aunque no estamos de acuerdo con todas las críticas realizadas por Martínez-García (2024), sí que se puede considerar que el artículo de M. Tourón et al. (2023) presenta algunas limitaciones derivadas de la falta de justificación teórica de los modelos propuestos, la toma de decisiones basadas en datos que pueden provocar sesgos por sobreajuste y el uso de factores reflectivos sin una consideración previa. Estos problemas no son exclusivos del artículo objeto de debate, sino que ejemplifican cuestiones que se encuentran ampliamente en la literatura, producto del proceso de *escritura automática* de artículos (según las palabras del propio Martínez-García (2024)) en el que muchas veces se cae debido a las presiones para intensificar la labor de difusión de resultados (Martínez-García & Martínez-Caro, 2009).

Análisis crítico de Martínez-García (2024)

Como señalábamos anteriormente, uno de los pilares de la crítica de Martínez-García (2024) al trabajo de M. Tourón et al. (2023) es la definición predeterminada de los factores latentes como reflectivos en todos los modelos definidos, sin contemplar la posibilidad de que sean formativos. Si bien estamos de acuerdo en el fondo con esta crítica, es importante señalar que la argumentación de Martínez-García (2024) se basa exclusivamente en cuestiones estadísticas y matemáticas, sin evidenciar la naturaleza formativa de estos factores en base a la redacción de sus ítems. De hecho, consideramos que Martínez-García (2024) presenta una visión reduccionista de los modelos reflectivos, cuando indica:

Los indicadores observables son reflectivos […] por lo que con un solo indicador se podría medir cada una de esas dimensiones. Si un investigador piensa que con un solo indicador es insuficiente para medir

una variable latente y que necesita más porque la dimensión latente es amplia […], lo que probablemente esté sucediendo es que esa batería de indicadores está midiendo variables latentes diferentes (Martínez-García, 2024)

Estamos en profundo desacuerdo con esta afirmación, bajo nuestro punto de vista existen fenómenos que presentan un amplio abanico de respuestas o conductas medibles que se espera que estén correlacionadas entre sí, y que en conjunto aportan matices, riqueza y una mayor exactitud en la medida del factor. Como ilustración de esta cuestión, solamente hace falta recordar el anterior *ejemplo del borracho*, en el que las posibles manifestaciones de la borrachera son muy variadas, por lo que, aunque estén correlacionadas, puede considerarse que si no se incluyeran todas las manifestaciones posibles se estaría reduciendo la precisión en la medida del factor.

Otra de las cuestiones clave que señala Martínez-García (2024), que ya hemos abordado brevemente más arriba, pero en la que nos queríamos detener más específicamente, es la referida al empleo de índices de ajuste absoluto e incremental complementarios al contraste chi-cuadrado. Martínez-García (2024) es categórico al afirmar que estos indicadores "comúnmente son usados para ajustar los modelos de forma que se evite el único test adecuado para detectar la mala especificación de un modelo (el test de la chi-cuadrado o, más correctamente, la familia de test de la chi-cuadrado)". A lo que añade que "cuando un modelo no pasa el test de la chi-cuadrado, los parámetros estimados no son interpretables, ya que una o varias de las relaciones de covarianza implicadas por el modelo no es apoyada por los datos empíricos, lo que de inmediato produce sesgo en las estimaciones" (Martínez-García, 2024). Podríamos aceptar este posicionamiento si el test chi-cuadrado fuera infalible, pero es bien conocido que esta prueba devuelve puntuaciones hinchadas (lo que supone una muy elevada tasa de rechazo de los modelos) ante condiciones de falta de normalidad multivariante (e.g., Curran et al., 1996; Hu et al., 1992), o que, como el propio Martínez-García (2024) admite, puede fallar en la detección de modelos mal especificados. Además, Martínez-García (2024) contrapone en su discusión la amplia literatura existente de autores de referencia en este campo, que "critican la prueba chi-cuadrado y recomiendan los índices de ajuste aproximados", con la crítica que realiza a estos índices Hayduk (2014), un investigador con niveles de difusión e impacto relativos en el ámbito de la investigación

Recibido: 01/06/2024 Aceptado: 01/06/2024

básica sobre modelización con estructuras de covarianzas claramente inferiores[1]. En este punto debemos preguntarnos si la relativa irrelevancia de Hayduk en comparación con investigadores de referencia como Jöreskog, Bentler, Steiger o Browne, se debe a que su trabajo no realiza una aportación significativa al conocimiento o a si su discurso resulta inconveniente para la publicabilidad de estudios que incluyen modelos de ecuaciones estructurales (SEM) y para la explotación comercial de los principales paquetes estadísticos que permiten su obtención.

Por último, al examinar en conjunto las críticas de Martínez-García (2024) apreciamos que no repara en la problemática que supone el estudio de los niveles de ajuste de los modelos con compuestos formativos. Es importante señalar que, como ya se ha mencionado, la definición de compuestos formativos no es formalmente posible en los modelos tradicionales basados en estructuras de covarianzas, o MEE. La teorización y validación de factores formativos es viable a través de métodos de mínimos cuadrados parciales (PLS-SEM), basados en las estructuras de varianzas de los datos. A diferencia de las técnicas basadas en estructuras de covarianzas, el PLS-SEM es más flexible en cuanto a los supuestos previos y el ajuste a distribuciones teóricas, siendo considerada una técnica de carácter exploratorio y confirmatorio que estudia relaciones de tipo predictivo-causal (Hair et al., 2022). De hecho, no es posible el empleo del test chi cuadrado para la verificación del del ajuste de los modelos PLS-SEM. Lo que es más, a pesar de que las herramientas actuales de PLS-SEM permiten estudiar el ajuste de los modelos a través de indicadores de ajuste aproximados (e.g., RMSEA, CFI, TLI), su empleo no está recomendado por buena parte de los investigadores de referencia en el ámbito (Hair et al., 2022). A esto se le suma, que en el caso de los compuestos formativos tampoco se aplican otros supuestos previos presentes en CFA como el análisis de las cargas de los ítems, el índice de fiabilidad compuesta o la varianza media extraída.

En suma, a pesar de que el análisis crítico de Martínez-García (2024) preconiza del uso del test chi cuadrado como única evidencia válida para la verificación del buen o mal ajuste del modelo, a la vez que defiende

[1] El trabajo fundamental de Hayduk en el que Martínez-García fundamenta su crítica (Hayduk, 2014) alcanza, en febrero de 2024, un total de 55 citas en Google Scholar, 40 citas en Web of Science, y 41 citas en Scopus.

Recibido: 01/06/2024 Aceptado: 01/06/2024

la existencia de dimensiones formativas en el modelo factorial de la escala GRS-2, no propone soluciones sobre cómo validar el ajuste de este modelo con dimensiones formativas sin el empleo de indicadores de ajuste aproximados.

Conclusiones

Podemos afirmar que el trabajo de M. Tourón et al. (2023), con sus fortalezas y sus limitaciones, es un buen ejemplo de cómo se están diseñando e implementando los procesos de validación estadística de escalas en Investigación Educativa. Así, las problemáticas abordadas en nuestro análisis crítico no son específicas de la propuesta de M. Tourón et al. (2023) y de Martínez-García (2024), sino que son cuestiones observadas de manera generalizada en la investigación psicométrica del ámbito de las Ciencias de la Educación.

En primer lugar, es habitual encontrar trabajos de validación psicométrica que no documentan la redacción completa de los ítems que conforman las escalas validadas (e.g., Fumero & Miguel, 2023; Quijada et al., 2020). Una de las razones fundamentales, como parece ocurrir en M. Tourón et al. (2023), es la existencia de derechos de propiedad intelectual sobre las escalas validadas. En otras ocasiones, nos encontramos con estudios publicados en un idioma diferente al de las escalas aplicadas en la muestra piloto. En estos casos es habitual que los ítems de las escalas se publiquen traducidas al idioma de la revista y no en el idioma original (e.g., Hernández Ramos et al., 2014; Quijada et al., 2020; Sánchez-Prieto et al., 2019), creando sesgos potenciales y dificultando el correcto uso de las mismas. A pesar de que estas prácticas son poco recomendables, están tan asentadas en el ámbito de las Ciencias Sociales y de la Educación que en muchos casos no son detectadas por editores, revisores o lectores especializados en la materia y son asumidas por los propios científicos como una buena praxis. Esta cuestión queda perfectamente ilustrada en las críticas de Martínez-García (2024), que más allá de no mencionar esta omisión, pone encima de la mesa la naturaleza formativa de las dimensiones sin conocer la lectura de sus ítems.

Otra cuestión importante es el desarrollo de un enfoque excesivamente empirista en los procesos de validación estadística de escalas, obviando en gran medida el modelo teórico desde el que ha sido construida la escala

Recibido: 01/06/2024 Aceptado: 01/06/2024

original. Esta perspectiva basada fundamentalmente en la búsqueda del modelo empírico con mejor ajuste puede llevar a soluciones con poco sentido teórico y práctico. Un buen ejemplo de este tipo de prácticas son trabajos que comparan el ajuste de varios modelos empíricamente plausibles sin justificar su relevancia teórica (e.g., Rodríguez Conde et al., 2012; Thomas et al., 2019), estudios que liberan parámetros de correlación entre los errores de algunos ítems basándose exclusivamente en los índices de modificación obtenidos (e.g., Thomas et al., 2019; J. Tourón et al., 2018), o investigaciones que intercambian los ítems entre las dimensiones disponibles, o incluso crean dimensiones nuevas, sin tener en cuenta los modelos teóricos y de indicadores a partir de los que se diseñó la escala (e.g., J. Tourón et al., 2018). Teniendo en cuenta que M. Tourón et al. (2023) no reparan en esta problemática en su análisis de limitaciones, y que Martínez-García (2024) centra sus críticas y realiza propuestas únicamente en relación a la identificación y ajuste de los modelos comparados, nuestra hipótesis es que los investigadores educativos hemos aprehendido y asumido este proceder. Así, nos preguntamos si es necesario reorientar los estudios de validación estadística en investigación educativa: ¿deberíamos sacrificar bondad de ajuste en los modelos (con la consecuente reducción de la publicabilidad de nuestros trabajos) a cambio de que éstos sean más coherentes y explicativos de la realidad? A nuestro juicio, debemos fomentar esta práctica integrando en nuestros procesos de diseño y validación de escalas equipos multidisciplinares que incluyan tanto expertos en los constructos abordados como especialistas en psicometría.

Aunque no profundiza en ello, Martínez-García (2024) destaca la *escritura automática* como una problemática muy existente en Ciencias Sociales. Tanto la falta de un conocimiento estadístico profundo, como asumir una postura acrítica sobre las implicaciones de los procedimientos metodológicos implementados, puede llevar al investigador educativo a la búsqueda de 'recetas' y la aplicación de protocolos de actuación rígidos en la validación de escalas. Este proceder orienta hacia un enfoque puramente empirista, alejado de la necesaria reflexión sobre las implicaciones teóricas y prácticas de los modelos estudiados. Bajo nuestro punto de vista, la aportación efectiva que realizan estos trabajos al avance y la mejora educativa es limitada.

En relación a la escritura automática, y debido fundamentalmente a la presión a la que se ven sometidos los académicos para aumentar su

producción científica, se observa en los últimos años la generalización de prácticas de *salami slicing* (Norman & Griffiths, 2008; Šupak Smolčić, 2013). Esta práctica éticamente dudosa consiste en separar, o trocear, una investigación con sentido unitario en varios informes de resultados, de modo que sea posible la publicación de varios trabajos diferentes. Lamentablemente, el *salami slicing* ha colonizado todas las áreas de la investigación científica, y estudios psicométricos desarrollados en investigación educativa como el aquí analizado no son una excepción (M. Tourón et al., 2023, 2024).

Por último, consideramos fundamental resaltar la falta de conciencia entre los investigadores educativos sobre la existencia de factores formativos, en comparación con los bien conocidos factores reflectivos. Esta falta de conocimiento lleva a que, en la mayor parte de los casos, se establezca un modelo reflectivo por defecto para la validación de escalas sin una reflexión previa sobre el modelo más apropiado. Dados los problemas asociados a tratar de validar factores formativos como reflectivos (Hair et al., 2011), los expertos en la materia debemos redoblar esfuerzos en difundir este tipo de modelos y democratizar su empleo en la investigación psicométrica.

En conclusión, el debate abierto en torno al artículo de M. Tourón et al. (2023) y la posterior crítica de Martínez-García (2024) ofrece la posibilidad de analizar una serie de prácticas naturalizadas en el desarrollo de estudios psicométricos en el ámbito educativo que limitan el desarrollo del campo. A nuestro entender, la sobreabundancia de estudios de esta naturaleza hace necesario dar un salto de calidad para que aumente su relevancia e implicaciones. Este salto de calidad pasa por ampliar el foco de las investigaciones que en la actualidad están excesivamente centradas en la reproducción casi mecanizada de una serie de procesos metodológicos y el cumplimiento de índices de ajuste dejando en segundo plano las implicaciones teóricas y prácticas de los modelos ajustados, que debería ser el objetivo principal de este tipo de estudios.

Referencias bibliográficas

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (Second edition). Guilford Publications.

Chin, W. W. (1998). Commentary: Issues and opinion on structural equation modeling. *MIS Quarterly*, *22*(1), vii–xvi.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2.ª ed.). Routledge. https://doi.org/10.4324/9780203771587

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29.

Educación XX1 (2024). *Normas para la presentación de originales*. https://revistas.uned.es/index.php/educacionXX1/libraryFiles/downloadPublic/170

Fumero, A., & Miguel, A. de. (2023). Validación de la versión española del NEO-FFI-30. *Análisis y Modificación de Conducta*, *49*(179). https://doi.org/10.33776/amc.v49i179.7325

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research: Univariate and multivariate applications*. Routledge Academic.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (3ª Ed.). Sage.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hayduk, L. A. (2014). 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

Hernández Ramos, J. P., Martínez Abad, F., García Peñalvo, F. J., Herrera García, M. E., & Rodríguez Conde, M. J. (2014). Teachers' attitude regarding the use of ICT. A factor reliability and validity study. *Computers in Human Behavior*, *31*, 509-516. https://doi.org/10.1016/j.chb.2013.04.039

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504. https://doi.org/10.1037/0033-2909.111.3.490

Martínez-García, J.A. (2024, en prensa). 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). *Revista de Educación*.

Martínez-García, J. A., & Martínez-Caro, L. (2009). El análisis factorial confirmatorio y la validez de escalas en modelos causales. *Anales de Psicología*, *25*(2), Article 2.

Norman, I., & Griffiths, P. (2008). Duplicate publication and «salami slicing»: Ethical issues and practical solutions. *International Journal of Nursing Studies*, *45*(9), 1257-1260. https://doi.org/10.1016/j.ijnurstu.2008.07.003

Quijada, A., Ruiz, M. A., Huertas, J. A., & Alonso-Tapia, J. (2020). Desarrollo y validación del Cuestionario de Clima Escolar para Profesores de Secundaria y Bachillerato (CES-PSB). *Anales de Psicología*, *36*(1), 155-165. https://doi.org/10.6018/analesps.36.1.341001

Rodríguez Conde, M. J., Olmos Migueláñez, S., & Martínez Abad, F. (2012). Propiedades métricas y estructura dimensional de la adaptación española de una escala de evaluación de competencia informacional autopercibida (IL-HUMASS). *Revista de investigación educativa*, *30*(2), 347-365.

Sánchez-Prieto, J. C., Hernández-García, Á., García-Peñalvo, F. J., Chaparro-Peláez, J., & Olmos-Migueláñez, S. (2019). Break the walls! Second-Order barriers and the acceptance of mLearning by first-year pre-service teachers. *Computers in Human Behavior*, *95*, 158–167. https://doi.org/10.1016/j.chb.2019.01.019

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! Journal of Business Research, 69(10), 3998-4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "Best" Factor Structure and Moving Measurement Validation Forward: An Illustration. *Journal of Personality Assessment*, *100*(4), 345-362. https://doi.org/10.1080/00223891.2018.1449116

Šupak Smolčić, V. (2013). Salami publication: Definitions and examples. *Biochemia Medica*, *23*(3), 237-241. https://doi.org/10.11613/BM.2013.030

Thomas, H. J., Scott, J. G., Coates, J. M., & Connor, J. P. (2019). Development and validation of the Bullying and Cyberbullying Scale for Adolescents: A multi-dimensional measurement model. *British Journal of Educational Psychology*, *89*(1), 75-94. https://doi.org/10.1111/bjep.12223

Tourón, J., Martín, D., Navarro-Asencio, E., Pradas, S., & Iñigo, V. (2018). Validación de constructo de un instrumento para medir la competencia digital docente de los profesores (CDD). *Revista española de pedagogía*, *76*(269), 25-54. https://doi.org/10.22550/REP76-1-2018-02

Tourón, M., Tourón, J., & Navarro-Asencio, E. (2023). Validez de Constructo de la Escala de Detección de Alumnado con Altas Capacidades para Profesores de Educación Infantil, Gifted Rating Scales (GRS2-P), en una muestra española. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *29*(2), Article 2. https://doi.org/10.30827/relieve.v29i2.27787

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

Información de contacto: Fernando Martínez Abad, Instituto Universitario de Ciencias de la Educación. Universidad de Salamanca.Paseo Canalejas, 169. 37008, Salamanca. E-mail: fma@usal.es