https://doi.org/10.21865/RIDEP66.5.02

Prácticas Cuestionables en Estudios de Validez de Instrumentos de Medición Psicológica: Comunalidades y Unicidades de la Crisis de Replicabilidad en el Campo de la Psicometría

Questionable Practices in Validity Studies of Psychological Measurement Tests: Communalities and Unicities of the Replication Crisis in the Field of Psychometrics

David Paniagua[1], Iván Sánchez-Iglesias[2], Alejandro Miguel-Alvaro[3], Nieves Casas-Aragonez[4], Marta Evelia Aparicio-Garcia[5] y Raimundo Aguayo-Estremera[6]

Resumen

En la última década se ha estudiado la crisis de replicabilidad: tanto los procesos como las soluciones. Dicho estudio ha supuesto una reflexión sobre el modo de hacer ciencia en psicología. En este trabajo se pretende resumir el estado actual de la crisis en replicabilidad enfocándolo hacia los estudios de validación de instrumentos de medición en psicología. Se explorarán las características que hacen que los estudios de validez tengan unas prácticas cuestionables específicas: Las Prácticas Cuestionables en estudios de Validez (PCV). Distinguimos tres grandes bloques de PCV: Teoría y Diseño, Ejecución y Redacción. Es necesario que se exploren la incidencia de las PCV, que los autores tomen conciencia de que son una mala práctica y que revistas y entidades aúnen esfuerzos en reducir su aparición.

Palabras clave: crisis replicabilidad, practicas cuestionables de investigación, prácticas cuestionables en estudios de validez

Abstract

In the last decade, the replication crisis has been studied: both the processes and the solutions. That study has involved a reflection on the way of doing science in psychology. This paper aims to summarize the current state of the crisis in replicability, focusing it on test validation studies in psychology. Specific characteristics that make Questionable Practices in Validity studies will be explored (QPV). We distinguish three large blocks of QPV: Theory and Design, Execution and Writing. It is necessary that the incidence of QPV be explored, that the authors become aware that they are a bad practice, and that journals and entities join forces to reduce their occurrence.

Keywords: replication crisis, questionable research practices, questionable practices in validity studies

[1] MSc. Personal Docente Investigador. Facultad de Psicología, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: davidpan@ucm.es
[2] PhD. Personal Docente Investigador. Facultad de Psicología, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: i.sanchez@psi.ucm.es
[3] MSc. Contratado predoctoral FPU. Facultad de Psicología, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: alemigue@ucm.e
[4] MSc. Licenciada en Psicología. Facultad de Psicología, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: niecasas@ucm.es
[5]PhD. Personal Docente Investigador. Instituto de Estudios Feministas, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: aparicio@psi.ucm.es
[6]PhD. Personal Docente Investigador. Facultad de Psicología, Universidad Complutense de Madrid, España. Campus de Somosaguas, Ctra. de Húmera, s/n, 28223 Pozuelo de Alarcón, Madrid. Correo: raaguayo@ucm.es (Autor de correspondencia)

ISSN: 1135-3848 print /2183-6051online

Introducción

Diversas líneas de investigación abordan la crisis de replicabilidad donde se valoran las causas contextuales e individuales. Por ejemplo, DuBois et al. (2013) se analiza la presión por publicar para conseguir una beca, o Ioannidis (2006) explora la dificultad de publicar resultados nulos. También abordan las conductas concretas relacionadas con la crisis de replicabilidad. Por ejemplo, las llamadas Prácticas Cuestionables en Investigación (PCI; Fiedler & Schwarz, 2016; John et al., 2012) y las posibles soluciones (e.g., García-Garzón et al., 2018; Blanco et al., 2017). Aunque quizá alguien pueda verla con un aura de oscuridad, la crisis de replicabilidad puede entenderse como un ejercicio de reflexión y autocorrección (Nelson et al., 2018). Como sugieren Asendorpf et al. (2013) y Lilienfeld y Waldman (2017), tomemos esta oportunidad para hacer una pausa, tomar aire y reconsiderar los estándares de calidad sobre los que trabajamos en nuestra área.

Aunque la bibliografía en el campo de la Psicología es amplia, son pocos los esfuerzos que se han dedicado a valorar la presencia, impacto y soluciones de la crisis de replicabilidad en el área de la Psicometría. A lo largo de este escrito se resumen los aspectos más destacados de la crisis de replicabilidad poniendo énfasis en las evidencias de prácticas cuestionables en estudios de psicología, así como en las potenciales soluciones y factores protectores. Después, se exploran los aspectos comunes (comunalidades) y particulares (unicidades) de la crisis de replicabilidad en el ámbito de la medición psicológica, en concreto, los estudios de validez. Se recogen evidencias de replicabilidad en el área de la psicometría y, finalmente, se describen algunas de las prácticas cuestionables al realizar estudios que tratan de obtener evidencias de validez.

Crisis de Replicabilidad

El problema de las malas conductas en investigación no es nuevo (Zuckerman, 1977, 1984), sin embargo, la denominada crisis de replicabilidad hizo saltar las alarmas en la investigación en psicología dentro del ámbito de las ciencias sociales y de la salud (Banks et al., 2016). Este término hace referencia a las dificultades para repetir estudios empíricos encontrando los mismos resultados que los publicados previamente en el campo de investigación (Heilmayr, 2021). Esto debería ser algo fundamental para la confirmación y refutación de los resultados de investigaciones que conforman nuestro conocimiento creíble (Ryan & Tipu, 2022; Nosek et al., 2022).

La preocupación generada por esta crisis ha llevado a numerosos intentos de replicación a gran escala, algunos de ellos muy publicitados, como el famoso estudio de la *Open Science Collaboration*, publicado en 2015, que concluye que solo el 39% de los estudios publicados eran replicables (Alister et al., 2021; Pearson et al., 2022). Así, la credibilidad de la investigación en este ámbito decae al encontrarnos en un contexto en el que se ha mostrado que la mayoría de los resultados de las investigaciones son falsos (Ioannidis, 2005). Una tasa tan alta es difícilmente justificable, a pesar de poder asumir un margen de error por multitud de causas y es, probablemente, indicador de un problema general en la disciplina (Alister et al., 2021).

¿A qué ha vinculado la literatura esta alta tasa? Entre otros factores, a la falta de entrenamiento de investigadores, el desarrollo de teorías ilusorias, uso incorrecto de las herramientas de medida o métricas en investigación, y debilidad metodológica y analítica (Aguayo et al., 2018; Ryan & Tipu, 2022). Entre estos factores, varios autores (v.g., Chambers, 2017) han enfatizado el papel de las llamadas Prácticas Cuestionables en Investigación (PCI).

Prácticas Cuestionables en la Investigación

Bajo este nombre se recogen prácticas o conductas no éticas aplicadas en cualquiera de los procesos de una investigación de forma intencionada en pos de conseguir resultados (v.g., publicar un artículo) en detrimento de la calidad (Fiedler & Schwarz, 2016; John et al., 2012). Son los llamados grados de libertad en la investigación que actúan alineados con los intereses personales del investigador y suelen estar incentivados por factores externos como la presión por publicar (Garfield, 1996; van Dijk et al., 2014), la financiación y políticas éticas débiles o con una

Tabla 1. PCI según el listado de John et al. (2012)

1. En un artículo, no informar de todas las medidas dependientes de un estudio.
2. Decidir si se recogen más datos después de comprobar si los resultados son significativos.
3. En un artículo, no informar de todas las condiciones de un estudio.
4. Dejar de recoger datos antes de lo previsto porque se ha encontrado el resultado que se buscaba.
5. En un artículo, "redondear" un valor *p* (por ejemplo, informar de que un valor *p* de .054 es inferior a .05).
6. En un artículo, informar selectivamente de los estudios que "funcionaron".
7. Decidir si se excluyen los datos después de analizar el impacto de hacerlo en los resultado.
8. En un artículo, informar de un hallazgo inesperado como si se hubiera predicho desde el principio.
9. En un artículo, afirmar que los resultados no se ven afectados por las variables demográficas (v.g., el género) cuando en realidad no se está seguro (o se sabe que sí).
10. Falsificación de datos.

implementación laxa en revistas científicas (Saunders & Savulescu, 2007).

Se han elaborado diferentes listados de PCI con el objetivo de poder estudiar su incidencia en investigaciones. Se ha propuesto, por ejemplo,una herramienta que contiene 10 PCI (ver Tabla 1) en las que los investigadores pueden incurrir al realizar una investigación en psicología utilizada en estudios con psicólogos investigadores de Brasil (Rabelo et al., 2019), Italia (Agnoli et al., 2017) y Estados Unidos (John et al., 2012) donde indican la existencia de PCI en este colectivo. Específicamente, en el estudio americano seis de las diez PCI del referido listado fueron empleadas por más de un 25% de de los encuestados y, de esas seis, dos fueron reportadas por más de la mitad de los encuestados ("No informar de todas las medidas dependientes de un estudio en el artículo" y "Decidir si se recogen más datos después de comprobar si los resultados son significativos"). El estudio italiano obtuvo resultados similares al estadounidense mientras que los psicólogos brasileños indicaron una menor tendencia a incurrir en esas dos PCI en comparación con sus homólogos italianos y norteamericanos, pero indicaron una mayor tendencia a incurrir en otras dos PCI ("informar selectivamente de los estudios que *funcionaban*" y "no informar de todas las condiciones de un estudio"). No obstante, estas prácticas no son específicas del campo psicológico sino que se extienden a todas las áreas de investigación y en todos los estratos de la carrera investigadora (e.g., Hofmann et al., 2013). Especialmente relevante en este sentido es poner el foco en la reciente revisión sistemática y meta-análisis de Xie et al., 2021 que aborda la prevalencia de las PCI de manera global hallando que un 12,5% de los investigadores informa haber cometido al menos una PCI y el 39,7% afirma ser consciente de que algún compañero la ha cometido.

Soluciones y Factores Protectores

Para prevenir y disminuir las PCI, y aligerar el impacto que tienen el resto de factores sobre la replicabilidad y reproducibilidad, se han resaltado algunos elementos, como la importancia de la metodología rigurosa y el empleo de herramientas de medición precisas (Flake & Fried, 2020; Vazire et al., 2022). Una de las grandes propuestas son las prácticas de ciencia abierta. En esta línea, García-Garzón et al. (2018) describen y valoran el impacto de la revisión por pares abierta (Iniciativa de Apertura de Revisores por Pares; Aleksic et al., 2014). Otras propuestas han sido las medidas de impacto alternativas, las directrices TOP para las revistas (*Transparence and Openesss Promotion*; *Center for Open Science*, 2017), el depósito de manuscritos científicos previos a su publicación (*preprints*) y el registro de artículos (Nosek & Lakens, 2014).

Por su parte, Waters et al. (2020) sugieren explorar las Prácticas de Calidad en las Publicaciones (PCP), dándole la importancia debida a los procesos de la preparación del manuscrito para su publicación (v.g., usando las guías CONSORT y PRISMA). Otros autores (v.g., Appelbaum et al., 2018; Shrout y Rodgers, 2018) pone el acento en potenciales iniciativas de las revistas para mejorar la calidad, fomentar prácticas de ciencia abierta (v.g., reconocimiento de autores), utilizar guías y estándares para realizar los distintos tipos de investigaciones. Asimismo, se ha recomendado que las

instituciones tomen partido, incluyendo metas y procedimientos para ayudar a los investigadores a superar las revisiones de calidad de sus trabajos y la adherencia a las antes mencionadas guías de calidad y estándares (Collins & Tabak, 2014; MacLeod et al., 2014).

Evidencias de Replicabilidad en la Investigación en Psicometría

Uno de los aspectos que mayor atención ha concitado en el área de la psicometría es el estudio de evidencias de validez, en concreto, sobre la estructura factorial de los instrumentos de medición (Cudeck et al., 2004). La relevancia concedida a esta área no parece desmesurada cuando se valora el papel que juega en el conjunto de evidencias que puede acumular una investigación. De este modo, buena parte de la investigación en psicología se apoya en el uso de instrumentos que deben contar suficiente respaldo psicométrico. Una vez que un estudio ha pasado el filtro de las revistas académicas, muchos de estos instrumentos se utilizan en decisiones que pueden conllevar consecuencias importantes. Así, en psicología forense, Schimmel y Van Koppen (2017) estudiaron un total de 1074 test psicológicos administrados por psicólogos forenses donde solo menos de la mitad (47%) de los tests fueron aceptados como con "suficiente calidad" por el comité oficial de evaluación de tests de Holanda. Otro caso similar se encuentra en Neal et al. (2019), donde solo el 40% de los test evaluados llegaban a un criterio aceptable de calidad psicométrica.

Los investigadores aplicados y revisores cuentan con multitud de herramientas a su disposición. Se pueden consultar manuales sobre psicometría (v.g., Abad et al., 2011; Martínez-Arias et al., 2006; Muñiz, 2018) y artículos que describen guías y recomendaciones (Ferrando et al., 2022; Toro et al., 2021). Los artículos de este tipo suelen publicarse en varias revistas de perfiles diferentes con actualizaciones periódicas y promueven aprendizajes sobre estadística, lo que puede reducir malas prácticas (Kuffner & Walker 2019). Teniendo en cuenta que las revistas son co-responsables del contenido que publican, no es de extrañar que se destinen tantos recursos para ofrecer guías de buenas prácticas y establecer una buena revisión por pares. Sin embargo, ¿la calidad de las investigaciones en psicometría alcanza los estándares propuestos por las propias revistas? Han sido varios los trabajos que han revisado la calidad de las publicaciones que tratan de obtener evidencias de validez factorial: Fabrigar et al. (1999) valoraron la calidad de los análisis factoriales exploratorios publicados por revistas de psicología concluyendo que la metodología informada en los artículos es "bastante deficitaria generalmente" (p. 295). Norris y Lecavalier (2010) concluyeron que "muchas investigaciones continúan usando metodología subóptima", que puede producir "soluciones (modelos) potencialmente distorsionados y sin sentido" (Ford et al., 1986, p. 307) y que puede afectar negativamente el desarrollo y mejora de teorías e instrumentos de medida (Bandalos & Gerstner, 2016; Fabrigar & Wegener, 2012; Haig, 2014; Henson & Roberts, 2006; Izquierdo et al., 2014; Lloret et al., 2017).

Como comentan Izquierdo et al. (2014): "desde que Ford, MacCallum y Tait (1986) publicaran el primer estudio de buenas prácticas en análisis factorial exploratorio, diferentes evaluaciones han puesto de manifiesto que algunos problemas persisten: el uso del método de extracción por componentes principales, el uso de un sólo método para extracción de factores (la extendida regla K1 de Kaiser) y la injustificada aplicación de modelos ortogonales (como Varimax)".

Esta problemática, expuesta en 1986, sigue vigente hoy en día, como se puede observar en los resultados de Izquierdo et al. (2014), Goretzko et al. (2021) y Ledesma et al. (2019); se puede obtener un resumen de por qué no son siempre correctos estos procedimientos en Ferrando (2022). En ellas se encuentran un uso del método de extracción por componentes principales en un 58% de los casos en ambas investigaciones, un uso del criterio K1 de Kaiser en un 20% de los casos y entre un 45% - 48% de los casos usaban métodos de rotación ortogonales. Además de problemas en las decisiones de llevar a cabo el AF, también encontraron falta de información reiterada en elementos clave para la reproducibilidad de los hallazgos como no informar de la matriz de correlaciones usada (74% - 86%), el software (24% - 36%) o información descriptiva (32,5%)

que permita valorar la adecuación del método de estimación a usar.

No solo hay problemas de calidad en las investigaciones donde se usan AFE, también existen en investigaciones que analizan la invarianza factorial (D'Urso et al., 2022). Las conclusiones de este estudio apuntan a la necesidad de mejorar la forma de comunicar los resultados, ya que esto es vital para la reproducibilidad y replicabilidad (Nosek et al., 2022). En este mismo estudio, recuerdan que Putnick y Bornstein (2016) sugirieron que se informe siempre de la muestra total y de cada grupo, la forma de afrontar la presencia de valores perdidos, la especificación del modelo, los índices de ajuste y la conclusión final sobre la invarianza. Esta información es importante no solo para los estudios psicométricos sino para cualquier estudio que asuma o requiera la invarianza de las medidas.

Además del estudio de la calidad de los artículos con aplicaciones de Análisis Factorial (FA), también se han analizado otros aspectos de la psicometría. En este terreno, más generalista, Flake y Fried (2020) acuñaron el concepto de Prácticas Cuestionables en Medición (PCM,) entendido como las decisiones de los investigadores que ponen en el punto de mira la validez de la medida . En la medición se incluyen la construcción y adaptación de instrumentos de evaluación psicológica, además de los métodos de escalamiento, corrección y comparación de puntuaciones. En este sentido, ya que a día de hoy no se conoce ningún constructo para el que exista un instrumento de medida natural, universal y rigurosamente validado, estas prácticas se verán afectadas por las Prácticas Cuestionables en Medición (Flake & Fried, 2020; Lilienfield & Strother, 2020).

Hemos recogido a lo largo de esta sección diversos aspectos que recogen evidencias acerca de las deficiencias metodológicas en las publicaciones del área de la psicometría. Citando de nuevo a Waters (2020), "en este escenario necesitamos hablar de elementos críticos en el proceso investigador incluyendo cómo son diseñados los estudios, cómo son generados los resultados, cómo se informa de ellos y cómo las revistas evalúan qué hace que un artículo tenga valor científico".

Prácticas Cuestionables en Estudios de Validez

Como se ha comentado hasta ahora, la crisis de replicabilidad ha sido ampliamente estudiada en los últimos años, ahondando en prácticas cuestionables, repercusiones sobre la calidad y factores protectores/soluciones. Diferentes artículos han resumido los distintos momentos del desarrollo de una investigación en psicología junto con los grados de libertad de la investigación que pueden llevar a realizar prácticas cuestionables. Basándonos en estos resúmenes de investigaciones en psicología vamos a valorar las *comunalidades* y *unicidades* de las partes del diseño, y posibles prácticas cuestionables, que se cometerían en el diseño prototípico de un estudio cuyo propósito es la obtención de evidencias de la validez interna de un test a partir de la técnica del Análisis Factorial (AF).

Picho y Artino (2016) comentan que las prácticas cuestionables se pueden dar antes de la investigación, con una revisión sesgada de la literatura o una mala elección de instrumentos de medida, o incluso después, por la falta de transparencia en los resultados. En este sentido, y tomando como base las PCI estudiadas en psicología, y lo comentado sobre las prácticas cuestionables en medición (PCM), mostramos a continuación los grados de libertad de los investigadores en publicaciones de validez de instrumentos de medida: las prácticas cuestionables en estudios de obtención de evidencias de validez, o dicho de forma más breve, prácticas cuestionables en validez (PCV).

Si el objetivo de un investigador fuera aumentar la probabilidad de publicar un estudio sobre obtención de evidencias de validez a costa de la calidad de la publicación, podría realizar diversas PCV, por ejemplo no informar de cargas factoriales cruzadas por debajo de ,40.. Hay que tener en cuenta que si el modelo no tiene un ajuste adecuado la probabilidad de ser publicado disminuye. Tampoco es habitual publicar estudios en los que el modelo hipotetizado no consigue un buen ajuste (similar al problema del *file-drawer*: estudios realizados con resultados no significativos que nunca se ven publicados); ni estudios para reproducir los análisis con modelos que obtuvieron buen ajuste y fueron publicados.

Tabla 2. Prácticas Cuestionables en estudios de Validez

T1. Generar ítems redundantes para mejorar la homogeneidad de la escala.
T2. Modificar la teoría para que se adecue al modelo inesperado que tiene buen ajuste.
T3. Decidir si se recogen o dejan de recoger datos después de comprobar el ajuste del modelo.
E1. Decidir si se excluyen datos después de analizar el impacto de hacerlo en el ajuste del modelo.
E2. Falsificar los datos o los valores de los resultados.
E3. Probar distintos métodos de estimación, rotación u otros elementos en contra de las recomendaciones metodológicas.
E4. Probar distintas submuestras de los datos recogidos (para AFE o AFC) y utilizar aquellas que mejor ajuste ofrezcan.
E5. Reespecificar el modelo utilizando índices de modificación sin informar de ellos.
E6. Hacer reespecificaciones o modificaciones de modelos no previstos teóricamente solo para mejorar el ajuste.
R1. No informar de todos los ítems puestos a prueba, solo de los seleccionados para el modelo que consigue el mejor ajuste.
R2. No informar de todos los modelos puestos a prueba mediante AF, solo del que concuerda con la teoría de partida o del que consigue el mejor ajuste.
R3. Redondear los pesos factoriales e índices de ajuste para que superen un punto de corte.
R4. Informar selectivamente de las especificaciones del modelo.
R5. Concluir que la escala está validada, omitiendo las limitaciones para generalizar los resultados.

Nota. T=Teoría y diseño; E=Ejecución; R=Redacción.

Hemos clasificado las PCV según la fase del proceso de investigación en tres grandes categorías: teoría y diseño, ejecución y redacción (ver Tabla 2).

Considerando los aspectos teóricos y de diseño de la escala se pueden incurrir en diversas PCV: (a) diseñar ítems que favorezcan ciertas estructuras o aumenten la consistencia interna (al que podemos referirnos con *λ-hacking* por su similitud con el *p-hacking*: tendencia a "redondear" el nivel crítico a ,05 o menos para justificar la significación en un contraste); (b) modificar la teoría para que se asemeje al modelo con mejor ajuste (parecido al *HARKing:* supone, por ejemplo, realizar primero el análisis factorial y en función del reparto de ítems y número de factores formular una teoría al respecto); y (c) algunas prácticas referidas al proceso de muestreo: ampliar, detener o seleccionar una muestra concreta que haga que el ajuste del modelo sea óptimo. Los dobletes o tripletes de ítems que no amplían contenido teórico pueden cumplir otras funciones, como aumentar la homogeneidad (consistencia interna) de la escala y generar factores espurios. Además, se puede trabajar con un mayor número de ítems del comunicado en el informe y seleccionar aquellos que mejor se ajusten a nuestras expectativas. Si la intención es establecer dos factores para mantener una hipótesis, se pueden generar los ítems de cada factor excesivamente parecidos entre ellos sin atender al contenido del constructo. Como sucede en los estudios de psicología en general, se podrían generar hipótesis a posteriori: redactar el modelo hipotetizado después de haber observado el ajuste. Otra opción para conseguir que los datos se ajusten a lo esperado es manipular los procesos de muestreo. Todas estas prácticas capitalizan el azar, ya que adecúan la teoría propuesta a lo obtenido en una muestra, lo que puede atentar directamente con las evidencias de validez externa. Es importante informar con detalle de estos detalles que se llevan a cabo en el muestreo como, por ejemplo, si la administración es online o en papel (Freiberg-Hoffmann & Romero-Medina, 2021).

En el proceso de ejecución las PCV se centran sobre todo en la toma de distintas decisiones sobre los análisis (v.g., factorial exploratorio, confirmatorio, invarianza). Se pueden cometer PCV como (a) excluir análisis estadísticos cuando no favorecen lo esperado (similar al *cherry-picking*); (b) falsificar los datos deliberadamente para mejorar el ajuste; (c) permutar todas las posibles combinaciones de métodos de estimación, rotaciones, número de factores a extraer y softwares en contra de las recomendaciones metodológicas, por ejemplo, usar la matriz de correlaciones de Pearson y el método de máxima verosimilitud con una escala de cuatro puntos que viola el supuesto de normalidad (parecido al *p-hacking*); (d) probar distintas submuestras de los datos recogidos y usar aquellas que ofrezcan mejores resultados, por

ejemplo, hacer el AFC la submuestra destinada al AFE o viceversa; (e) modificar el modelo en base a los índices de modificación sin informar de ello; y (f) hacer reespecificaciones, modificaciones o selección de modelos no previstos teóricamente y derivados de los índices de modificación de la muestra para mejorar el desajuste, sin mencionar la capitalización del azar (de nuevo, similar al *p-hacking;* como ejemplos: utilizar índices de modificación para correlacionar errores o permitir cargas cruzadas no esperadas teóricamente. Todo ello sin clarificarlo o sin tener en cuenta la capitalización del azar).

Finalmente, tan importante es usar los métodos y técnicas adecuadas como informar sobre ellas de forma precisa y exhaustiva. La transparencia en el informe de investigación es la única manera de que el lector pueda valorar la veracidad de los hallazgos en su conjunto. Por ello las PCV incluyen elementos relativos a la redacción y presentación de resultados en el informe de investigación. Algunas de las que se pueden cometer son: (a) informar solamente de los ítems que forman la estructura factorial que consigue un buen ajuste, omitiendo aquellos ítems que inicialmente formaban parte de ella pero que, por problemas de ajuste del modelo, se eliminaron (que podríamos llamar *item-picking* por la similitud con el *cherry-picking*); (b) omitir información sobre los AF realizados e informar solo del que concuerda con el modelo hipotetizado o del que consigue el mejor ajuste; (c) redondear o maquillar los valores los pesos factoriales o de los índices de ajuste (v.g., RMSEA, CFI, GFI) para que superen un determinado umbral; (d) omitir información importante de las especificaciones del modelo, por ejemplo, no comunicar la especificación de los errores correlacionados, o de la liberación de parámetros en estudios de invarianza factorial; y (e) concluir que la escala está validada, omitiendo las limitaciones del estudio para generalizar los resultados a otras aplicaciones con distinta población o contexto.

Conclusión

Como sugiere Bakker et al. (2012), la ciencia puede llegar a percibirse como un juego, en el que los jugadores (investigadores), formando equipos a veces, tratan de ganar partidos (artículos de impacto y citas) y trofeos (premios y becas) jugando con las reglas establecidas (normas de publicación de artículos y evaluación de investigadores, formas de acceso y promoción en la carrera investigadora) que son administradas por los árbitros (revisores, editores de revistas, empresas editoriales y políticos). El clima actual de gratificación hace que la mayor parte de los beneficios para un investigador pasen por publicar mucho, cuanto más mejor; el conocido lema "*publish or perish*" lo refleja bien. Esto no solo afecta al investigador y su grupo sino también al resto de elementos que conforman la comunidad científica (Begley & Ioannidis, 2015; Dunn et al., 2019). Este clima contiene elementos explícitos, como los que se pueden deducir de las palabras de Bakker et al. (2012), pero también implícitos, por ejemplo, una cultura competitiva, basada en el logro individual, donde se promueve más la innovación frente a la acumulación de evidencias de validez (Begley & Ioannidis, 2015; LeVeque et al., 2012; MacLeod et al., 2014).

Teniendo en cuenta que los autores de las publicaciones que tratan la validez de los instrumentos de medición psicológica no están en un contexto distinto, no es descabellado suponer que las PCI estudiadas en la literatura en psicología se repiten en el área de la medición. Como un estudio de validez tiene diferencias respecto de un estudio de psicología general y de otros que tratan aspectos diferentes de la medición, este trabajo ha intentado especificar las características idiosincráticas que hacen que los autores de estudios de validez lleguen a cometer PCI, lo que hemos llamado prácticas cuestionables en estudios que buscan evidencias de validez o, más breve, prácticas cuestionables en estudios de validez (PCV).

De todo lo expuesto en este trabajo se puede concluir que los investigadores tienen bastantes grados de libertad al hacer análisis factorial (principal herramienta para obtener evidencias sobre la estructura de un test). Esto aumenta la posibilidad de obtener resultados equivocados y disminuir la reproducibilidad de estos estudios y la replicabilidad de los hallazgos, al tiempo que la credibilidad en el área. Por ello, abogamos por hacer una pausa y reflexionar acerca de las consecuencias de publicar estudios de validez donde abundan las prácticas cuestionables. Una

consecuencia destacada tiene relación con el número y variedad de instrumentos existentes. En ocasiones, los investigadores aplicados tienen a su disposición una amplia variedad de tests, escalas y cuestionarios que miden el mismo constructo. En otras, disponen de variedad de instrumentos que miden supuestamente constructos distintos, aunque no haya suficiente evidencia. Los problemas de replicabilidad y reproducibilidad que hemos mencionado en este trabajo pueden estar relacionados con este auge injustificado de instrumentos de medición. En psicología, los investigadores están principalmente recompensados por hacer investigación novedosa, el número de publicaciones producidas y otros hitos académicos que acentúan el valor puesto en la innovación frente a el desinterés por mejorar la reproducibilidad y replicabilidad (MacLeod et al., 2014). Futuras investigaciones deberían valorar empíricamente la presencia de las prácticas cuestionables en estudios de validez de acuerdo con la propuesta realizada en el presente trabajo, además de reflexionar sobre la naturaleza, implicaciones y soluciones a las prácticas cuestionables específicas en el campo de la validez de los instrumentos de medición.

También hay que tener en cuenta que las revistas y revisores cumplen un papel fundamental en la rectificación de este tipo de prácticas y son partícipes del trabajo que resulta publicado. Tan importante es planificar bien el diseño, seleccionar el método correcto, realizar los análisis precisos y redactar el informe de investigación de forma transparente como la toma de conciencia por parte de las revistas acerca de su responsabilidad en la calidad de la información publicada y el libre acceso a los materiales que la han producido. La pobreza informativa en los artículos publicados socava la evaluación crítica que se puede hacer. Los lectores necesitan información completa, clara y transparente para evaluar plenamente la investigación (Agha et al., 2016). Por ello, se recomienda a las revistas que refuercen la calidad de la información publicada, sobre todo ahora que las posibilidades de almacenamiento digital parecen casi ilimitadas.

Los investigadores necesitan ofrecer más detalles en sus secciones de métodos y resultados. Como indican Wigboldus y Dotsch (2016):

"sí, correremos el riesgo de molestar a nuestros lectores con «historias de aflicción» (ver Bem, 1987), alargando nuestros artículos o escribiendo adendas en línea que pueden no ser tan fáciles de leer para los lectores menos informados. Sin embargo, esto puede ser la única forma de evitar falsos positivos sin el riesgo de obtener demasiados falsos negativos."

Por ello, las revistas deberían favorecer espacios donde se recompense valorar la calidad de los estudios publicados. Esto puede gestionarse con prácticas de ciencia abierta donde se tenga acceso a *preprints* o *registered reports* que especifiquen al detalle el procedimiento y las decisiones llevadas a cabo (métodos de estimación, rotación, especificación del modelo…) antes de recoger la muestra. Además, las revistas en las que sea habitual publicar estudios de validez pueden revisar el estado de la calidad de sus publicaciones promoviendo números especiales en los que se revisen si en artículos publicados hay una buena adherencia a los estándares de calidad. En resumen, se trata de aunar esfuerzos, por parte de autores, revisores y editores en crear, adaptar, popularizar y asegurar que se siguen guías de calidad en las publicaciones relacionadas con la psicometría. Algunas de esas normas y estándares son la COSMIN (Gagnier et al., 2021), el decálogo de Ferrando et al. (2022) y las recomendaciones de Flake y Fried (2020). A este respecto, creemos que un aspecto fundamental consiste en incluir, al menos, a un experto en psicometría en todos los procesos de revisión por pares de estudios de validez.

Las instituciones también deben implicarse. Waters et al., (2020) recogen propuestas directas a las instituciones para que generen políticas y prácticas que presenten un conjunto de valores de calidad en las publicaciones, que sea ampliamente compartido, fuertemente arraigado y transversal a todos los niveles de la comunidad científica (LeVeque et al., 2012; Nosek et al., 2015).

Las publicaciones se han convertido en la base de un conflicto en la ciencia: el interés personal de los investigadores y el interés de la comunidad científica por avanzar en nuestros conocimientos. La razón estriba en que publicar no es sinónimo de verdad. En la ciencia actual lo que se recompensa es la publicación, no la publicación rigurosa o verdadera (Nosek, 2012).

Muchos investigadores incurren en PCI porque facilitan la posibilidad de conseguir más publicaciones, sean éstas más o menos rigurosas. No obstante, dichas prácticas tienen una naturaleza multicausal, por lo que se necesitan acciones en diversos frentes para prevenirlas y erradicarlas. La psicología y las personas sobre las que se utilizan los instrumentos de medición desarrollados en nuestro campo de conocimiento merecen una evaluación rigurosa y resultados con garantías suficientes. Si queremos que la calidad de los estudios de validez mejore debemos poner de nuestra parte autores, revisores, revistas e instituciones. Como dijeron Nosek et al. (2015), cambiar la cultura de investigación no es una tarea fácil, pero es una meta importante hacia la que debemos dirigirnos.

Referencias

Abad, F. J., Olea, J., Ponsoda, V., & García., C. (2011). Medición en Ciencias Sociales y de la Salud. Síntesis.

Agha, R. A., Lee, S. Y., Jeong, K. J. L., Fowler, A. J., & Orgill, D. P. (2016). Reporting quality of observational studies in plastic surgery needs improvement: A systematic review. *Annals of Plastic Surgery*, *76*(5), 585-589. https://doi.org/doi:10.1097/SAP.0000000000000419

Agnoli, F., Wicherts, J., Veldkamp, C., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE 12*(3). https://doi.org/10.1371/journal.pone.0172792

Aguayo, R. (2018). La investigación en el síndrome de burnout: Reflexión crítica desde una perspectiva metodológica. *Apuntes de Psicología*, 36(1-2), 93-100. https://doi.org/10.55414/ap.v36i1-2.715

Aleksic, J., Alexa, A., Attwood, T. K., Hong, N. C., Dahlö, M., Davey, R., Dinkel, H., Förstner, K. U., Grigorov, I., Hériché, J.-K., Lahti, L., MacLean, D., Markie, M. L., Molloy, J., Schneider, M. V., Scott, C., Smith-Unna, R., & Vieira, B. M. (2014). An open science peer review oath. *F1000Research, 3*. https://doi.org/10.12688/f1000research.5686.1

Alister, M., Vickers-Jones, R., Sewell, D. K., & Ballard, T. (2021). How do we choose our giants? Perceptions of replicability in psychological science. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211018199

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report, 73, American Psychologist3–25. https://doi.org/10.1037/amp0000191.

Asendorpf, J., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J., Fiedler, K., Fiedler, S., Funder, D., Kliegl, R., Nosek, B., Perugini, M., Roberts, B., Schmitt, M., Van Aken, M., Weber, H., Wicherts, J. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108-119. https://doi.org/10.1002/per.1919

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. Perspectives on Psychological Science, 7(6), 543-554. https://doi.org/10.1177/1745691612459060

Bandalos, D., & Gerstner, J. J. (2016). Using factor analysis in test construction. En K. Schweizer, K. y DiStefano, C. (Eds.), *Principles and methods of test construction: Standards and recent advances* (pp. 26-51). Hogrefe Publishing.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323-338. https://doi.org/10.1007/s10869-016-9456-7

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (Eds.), The complete academic: A practical guide for the beginning social scientist (pp. 171-201). Random House.

Begley, G. C., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. Circulation Research, 116-126.

Blanco, F., Perales, J.C., & Vadillo,M.A. (2017). Pot la psicologia rescatar-se a si mateixa?

Incentius, biaix i replicabilitat. Anuari de psicologia de la Societat Valenciana de Psicologia, *18*(2), 231-252.

Chambers, C. (2017). The seven deadly sins of psychology. In The Seven Deadly Sins of Psychology. Princeton University Press.

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. Nature, 505, 612-613.

Cristea, I. O., & Naudet, F. (2019). Increase value and reduce waste in research on psychological therapies. Behaviour Research and Therapy.

Cudeck, R. (2007). Factor analysis in the year 2004: Still spry at 100. In Factor Analysis at 100 (pp. 15-22). Routledge.

Dubois, J. M., Anderson, E. E., Chibnall, J., Carroll, K., Gibb, T., Ogbuka, C., & Rubbelke, T. (2013). Understanding research misconduct: A comparative analysis of 120 cases of professional wrongdoing. *Accountability in Research*, *20*(5-6), 320-338.

D'Urso, E. D., Maassen, E., van Assen, M. A., Nuijten, M. B., De Roover, K., & Wicherts, J. (2022). The dire disregard of measurement invariance testing in psychological science.

Dunn, B., O'Mahen, H., Wright, K., & Brown, G. (2019). A commentary on research rigour in clinical psychological science: How to avoid throwing out the innovation baby with the research credibility bath water in the depression field. Behaviour Research and Therapy.

Fabrigar, L., Wegener, D., MacCallum, R., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fabrigar, L., & Wegener, D. (2012). *Exploratory factor analysis.* Oxford University Press.

Ferrando, P., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the Factor Analysis of Test Items. *Psicothema 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45-52. https://doi.org/10.1177/1948550615612150

Flake, J., & Fried, E. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Freiberg-Hoffmann, A., & Romero-Medina, A. (2021). Approaches and Study Skills Inventory for Students: Comparación de las propiedades psicométricas entre las versiones de lápiz-papel y online en estudiantes universitarios. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica,* 4(64), 17-30. https://doi.org/10.21865/RIDEP61.4.11

Ford, J., MacCallum, R., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*(2), 291-314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

García-Garzón, E., Lecuona, O., & Carbajal, G. (2018). Estudios de replicación, pre-registros y ciencia abierta en Psicología. *Apuntes de Psicología, 36* (1-2), 75-83. https://doi.org/10.55414/ap.v36i1-2.713

Garfield, E. (1996). What is the primordial reference for the phrase 'publish or perish'. The Scientist, *10*(12), 11.

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, 40(7), 3510-3521. https://doi.org/10.1007/s12144-019-00300-2

Haig, B. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* MIT press.

Heilmayr, D. (2022). A course unit and presentation assignment to teach students about open science and replicability in psychology. *Scholarship of Teaching and Learning in Psychology.* https://doi.org/10.1037/stl0000324

Henson, R., & Roberts, J. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Hofmann, B., Myhr, A., & Holm, S. (2013). Scientific dishonesty - A nationwide survey of doctoral students in Norway. *BMC Med Ethics 14*, 3. https://doi.org/10.1186/1472-6939-14-3

Ioannidis, J. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Izquierdo, I., Olea, J., & Abad, F. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, *26*(3), 395-400. https://doi.org/10.7334/psicothema2013.349

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kuffner, T., & Walker, S. (2019). Why are p-values controversial?. *The American Statistician*, *73*(1), 1-3. https://doi.org/10.1080/00031305.2016.1277161

Ledesma, R., Ferrando, P., & Tosi, J. (2019). Uso del Análisis Factorial Exploratorio en RIDEP. Recomendaciones para Autores y Revisores. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, *52*(3), 173-180. https://doi.org/10.21865/RIDEP52.3.13

LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing. *Computing in Science & Engineering*, 13-17.

Lilienfeld, S., & Waldman, I. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions.* John Wiley & Sons.

Lilienfeld, S., & Strother, A. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology / Psychologie canadienne, 61*(4), 281-288. https://doi.org/10.1037/cap0000236

Lloret, S., Ferreres, A., Hernández, A., & Tomás, I. (2017). El análisis factorial exploratorio de los ítems: Análisis guiado según los datos empíricos y el software. *Anales de Psicología/Annals of Psychology, 33*(2), 417-432. https://doi.org/10.6018/analesps.33.2.270211

MacLeod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., et al. (2014). Biomedical research: Increasing value, reducing waste. The Lancet, 383(9912), 101-104. https://doi.org/10.1016/s0140-6736(13)62329-6.

Martínez-Arias, M. R., Hernández-Lloreda, M. J. y Hernández-Lloreda, M. V. (2006). Psicometría. Alianza Editorial.

Muñiz, J. (2018). Introducción a la Psicometría. Teoría Clásica y TRI. Pirámide.

Norris, M., & Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developmental Disorders, 40*(1), 8-20. https://doi.org/10.1007/s10803-009-0816-2

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results [Editorial]. *Social Psychology*, 45(3), 137-141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B., Hardwicke, T., Moshontz, H., Allard, A., Corker, K., Dreber, A., Fidler, F., Hilgard, J., Struhl, M., Nuijten, M., Rohrer, J., Romero, F., Scheel, A., Scherer, L., Schönbrodt, F., & Vazire, S. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology* 2022 73:1, 719-748

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Pearson, M., Schwebel, F., Richards, D., & Witkiewitz, K. (2022). Examining replicability in addictions research: How to assess and ways forward. *Psychology of Addictive Behaviors, 36*(3), 260-270. https://doi.org/10.1037/adb0000730

Picho, K., Maggio, L. A., & Artino, A. R. (2016). Science: The slow march of accumulating evidence. *Perspectives on Medical Education, 5*(6), 350-353. https://doi.org/10.1007/s40037-016-0305-1

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and

reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Rabelo, A., Farias, J., Sarmet, M., Joaquim, T., Hoersting, R., Victorino L., Modesto, J., & Pilati, R. (2019). Questionable research practices among Brazilian psychological researchers: Result from replication study and an international comparison. *International Journal of Psychology*, *55*(4), 674-683. https://doi.org/10.1002/ijop.12632

Ryan, J., & Tipu, S. (2022). Business and management research: Low instances of replication studies and a lack of author independence in replications. *Research Policy, 51*(1), 104408. https://doi.org/10.1016/j.respol.2021.104408

Saunders, R., & Savulescu, J. (2008). Research ethics and lessons from Hwanggate: What can we learn from the Korean cloning fraud?. *Journal of Medical Ethics, 34*(3), 214-221. http://doi.org/10.1136/jme.2007.023721

Schimmel, L. M. C., & Van Koppen, P. J. (2017). Verdachten testen: Testgebruik in de forensische psychologie [testing suspects: Use of tests in forensic psychology]. De Psycholoog, 52(10), 34-42. https://www.tijdschriftdepsycholoog.nl/wetenschap/verdachten-testen/

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487-510.

Simmons, J., Nelson, L., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*(2), 255-259. https://doi.org/10.1177/1745691617698146

Toro, R., Peña-Sarmiento, M., Avendaño-Prieto, B. L., Mejía-Vélez, S., & Bernal-Torres, A. (2022). Análisis empírico del Coeficiente Alfa de Cronbach según opciones de respuesta, muestra y observaciones atípicas. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliaçãol Psicológica*, *63*(2), 17-30. https://doi.org/10.21865/RIDEP63.2.02

Van Dijk, D., Manor, O., & Carey, L. (2014). Publication metrics and success on the academic job market. *Current Biology, 24*(11), R516-R517. https://doi.org/10.1016/j.cub.2014.04.039

Vanier, S., Schiavone, S., & Bottesini, J. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162-168. https://doi.org/10.1177/09637214211067779

Waters, A. M., LeBeau, R. T., Young, K. S., Dowell, T. L., & Ryan, K. M. (2020). Towards the enhancement of quality publication practices in clinical psychological science. *Behaviour Research and Therapy, 124*, 103499. https://doi.org/10.1016/j.brat.2019.103499

Wigboldus, D.H.J., Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika* 81, 27-36). https://doi.org/10.1007/s11336-015-9445-1

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics, 27*(4), 1-28. https://doi.org/10.1007/s11948-021-00314-9

Zuckerman, H. (1977). Deviant behavior and social control in science. In E. Savarin (Ed.), Deviance and social change (pp. 87-138). Sage.

Zuckerman, H. (1984). Norms and deviant behavior in science. *Science, Technology, & Human Values*, *9*(1), 7-13.