Estado de las prácticas científicas e investigación educativa. Posibles retos para la próxima década

State of scientific practices and educational research. Potential challenges for the next decade

DOI: 10.4438/1988-592X-RE-2017-381-386

Ángeles Blanco-Blanco
Universidad Complutense de Madrid

Resumen

El trabajo presenta una revisión del estado de las prácticas científicas actuales y de su impacto potencial en la calidad de la investigación educativa. Desde una concepción post-positivista de la investigación científica en educación, el problema se desarrolla en el contexto general del debate actual sobre la ciencia, sobre su fiabilidad, su robustez y su reproducibilidad. Teórica y conceptualmente el estudio adopta el enfoque propio de la meta-investigación. Desde el punto de vista metodológico, se lleva a cabo una revisión de la literatura sobre el tema que permita una reflexión fundamentada sobre el *status quo* de las prácticas científicas. Para ello se consideran algunos de los trabajos más relevantes y de mayor impacto publicados en los últimos años sobre meta-ciencia, en general y en el ámbito de las Ciencias de la Educación y del Comportamiento en particular. En primer lugar se caracteriza lo que se ha etiquetado como *crisis de la ciencia* y particularmente se trata de la presencia de sesgos y prácticas de investigación cuestionables en la investigación científica. A continuación se presentan algunos de los elementos correctores clave que se han propuesto para fortalecer y permitir un avance más efectivo de la empresa científica. Estos incluyen: alternativas y "nuevos" énfasis en el análisis estadístico de datos científicos; el impulso renovado a la replicación y la reproducibilidad; y los nuevos modos de producción, difusión y evaluación de la investigación ligados a la ciencia abierta. El artículo se cierra con algunas reflexiones relativas a posibles retos para la investigación educativa española en la próxima década. Las conclusiones se organizan en torno a cuatro ejes: el desarrollo de estudios meta-investigativos;

Recibido: 09-04-2017 Aceptado: 23-02-2018

la formación, información y sensibilización de los investigadores sobre prácticas de investigación cuestionables; la actualización de las políticas editoriales; y el papel de los financiadores y evaluadores de la producción científica.

Palabras clave: Prácticas de investigación cuestionables, Crisis de reproducibilidad, Ciencia abierta, Ecosistemas de la producción y la publicación científicas, Meta-investigación, Investigación educativa.

Abstract

This paper presents a review of the state of current scientific practices and their potential impact on the quality of educational research. From a post positivist conception of scientific research in education, the matter is addressed in the general context of the current debate about science, its reliability, robustness and reproducibility. Theoretically and conceptually, the study adopts a meta-research approach. From a methodological perspective, a review of the literature on the subject is carried out that allows a reasoned reflection on the *status quo* of scientific practices. This is done by considering some of the most relevant papers published in recent years on meta-science in general, and in the field of Education and Behavioral Sciences. In the first place, the so-called *crisis of science* is characterized, and particularly the presence of biases and questionable research practices in scientific research. Below are some of the key corrective elements proposed to strengthen and enable a more effective advancement of the scientific enterprise. These include: alternative and "new" emphasis on the statistical analysis of scientific data; renewed impetus to replication and reproducibility; and new modes of production, dissemination and assessment of research associated with open science. The article closes with some reflections regarding possible challenges for Spanish educational research in the next decade. The conclusions are organized around four axes: the development of meta-research studies; training, information and awareness-raising of researchers about questionable research practices; updating editorial policies; and the role of funders and evaluators of scientific production.

Key words: Questionable research practices, Reproducibility crisis, Open Science, Scientific ecosystem, Meta-research, Educational research.

La investigación empírica en educación está llamada a informar las prácticas y las intervenciones educativas. Pero para poder contribuir de un modo efectivo a la toma de decisiones una condición necesaria, aunque ciertamente no suficiente, es que la investigación educativa sea

Recibido: 09-04-2017 Aceptado: 23-02-2018

científicamente robusta, fiable y válida. En este estudio proponemos una reflexión sistemática sobre dicha condición. Presentamos una revisión de las prácticas científicas actuales para analizar su impacto potencial sobre la calidad y el rigor de la investigación educativa. El análisis se realiza desde una concepción post-positivista de la investigación científica en educación y en el contexto general del debate actual sobre la ciencia, sobre su fiabilidad, su robustez y su reproducibilidad. Desde el punto de vista teórico-conceptual el estudio adopta el enfoque propio de la meta-investigación. Desde el punto de vista metodológico, se lleva a cabo una revisión de la literatura sobre el tema que permita una reflexión fundamentada acerca del *status quo* de las prácticas científicas actuales. Para ello se consideran algunos de los trabajos más relevantes y de mayor impacto publicados en los últimos años sobre meta-ciencia, en general y en el ámbito de las Ciencias de la Educación y del Comportamiento en particular.

Crisis de la ciencia y meta-investigación

En los últimos años se ha asistido en un amplia variedad de campos disciplinares a una fuerte crisis de confianza en lo que podríamos llamar el *status quo* de las prácticas científicas. Se ha hablado de *crisis de la ciencia* y, particularmente, de una *crisis de reproducibilidad* (Munafò et al., 2017). Un buen indicador del estado de cosas en este sentido lo representan las editoriales simultáneas publicadas por las revistas *Science* y *Nature* en 2014 bajo un elocuente título: *Journals unite for reproducibility* (Nature, 2014; McNutt, 2014). La editorial da cuenta de un acuerdo conjunto para adoptar nuevos principios y directrices en la publicación de investigación clínica. Pero lo esencial y que queremos destacar aquí es que se trataba así de reforzar las iniciativas que en los últimos años se venían tomando como respuesta a la crisis generada por las evidencias de falta de reproducibilidad, fraudes y malas prácticas que venían minando la confianza en la ciencia no sólo de los profesionales, sino de la opinión pública en general.

La preocupación y el estado de opinión en la comunidad científica puede ser ilustrado con algunos datos obtenidos en una encuesta reciente publicada en *Nature* (Baker, 2016). El 90% de los 1576 investigadores encuestados cree que efectivamente se asiste a una crisis

Recibido: 09-04-2017 Aceptado: 23-02-2018

de reproducibilidad y el 52% la califican como significativa. Las causas más destacadas incluyen prácticas de investigación cuestionables (p.e. informe selectivo o *p-hacking*), la presión por publicar y cuestiones vinculadas al uso del aparato estadístico y el diseño de investigación. Esta última dimensión, la estadístico-metodológica, encabeza la lista de soluciones señaladas por los encuestados, particularmente lo referido a la mejora en la formación y en la supervisión estadístico-metodológica. De hecho se ha apuntado que la *crisis* podría estar en buena parte derivada del mal uso del apartado estadístico en la investigación científica (Goodman, 2016; Peng, 2015).

Las dificultades encontradas para reproducir la investigación publicada deben principalmente verse como un síntoma, como el resultado de un problema de fondo complejo vinculado a diversos factores que expresan el escenario y los parámetros bajo los cuales se hace y difunde la ciencia. Como recuerdan Makel y Plucker (2014), que un resultado no sea replicable no le convierte en falso. Y obviamente un resultado replicable no es por ello verdadero. Pero ciertamente la replicabilidad es una pieza angular del sistema de la ciencia y la discusión en torno a la misma parece haber puesto un foco de atención necesario sobre el rigor y la transparencia de las prácticas científicas[1].

Donde posiblemente más fuerte y turbulenta ha sido la crisis en los últimos años es en el ámbito de la investigación en Psicología (Ledgerwood, 2014), un campo relativamente afín al educativo. La detección de fraudes a cargo de reputados investigadores, la publicación de dudosos y controvertidos descubrimientos y la evidencia de prácticas cuestionables extendió la sospecha de que quizá no todo funciona bien tras las elevadas tasas de éxito con la que los investigadores predicen los resultados que hallan en sus propios estudios. Así Simmons, Nelson & Simonsohn (2011) sugirieron que la hiper-precisión en la predicción (p.e. rechazo de la hipótesis nula) es probablemente debida a varios factores, que incluyen: recogida de datos hasta que se encuentra el resultado

(1) Efectivamente la replicación, entendida como la repetición de estudios previos a cargo de investigadores independientes, es la vía esencial para acumular evidencia científica a favor o en contra de las hipótesis bajo estudio. La replicabilidad, por tanto, está relacionada con la probabilidad de que un estudio independiente produzca resultados consistentes con el estudio original. El concepto de reproducibilidad es una variación reciente de este concepto (Peng, 2015). Se refiere a un estándar de mínimos especialmente referido a la investigación que implica computación masiva de datos. Se define como la posibilidad de recalcular los resultados originales dadas la base de datos y las especificaciones analíticas originales (p.e. código y documentación).

Recibido: 09-04-2017 Aceptado: 23-02-2018

deseado, no informar de ensayos fallidos, eliminación a posteriori de observaciones o variables que no apoyan las hipótesis planteadas. Es decir, un elenco más o menos amplio de prácticas de investigación cuestionables.

Todo lo anterior ha promovido un amplio y vivo debate en el ámbito de la investigación psicológica en los últimos años, del que no daremos cuenta de modo exhaustivo pero que puede ser bien ilustrado a partir de al menos dos indicadores:

- La publicación de numerosas secciones especiales y trabajos de discusión sobre prácticas de investigación cuestionables, replicabilidad y reproducibilidad en reconocidas revistas, tales como *Perspectives on Psychological Science* (2012, 2014) o *Psykometrica* (Sijtsma, 2016; Sijtsma, Veldkamp & Wicherts, 2016; Waldman & Lilienfeld, 2015; Wigboldus & Dotch, 2015).
- El impulso de ambiciosos programas de investigación a gran escala orientados específicamente a replicar resultados de la investigación psicológica publicada, adoptando fundamentalmente nuevas formas de trabajo colaborativo entre equipos de investigación independientes y consorcios de investigación (véase en este sentido Open Science Collaboration, 2015 o Schweinsberg et al., 2016).

En el ámbito de la investigación educativa nos parece que el eco de esta crisis quizá ha sido menor. Ello no quiere decir que posiblemente no se registren problemas similares. Creemos más bien que aparentemente no se ha reunido evidencia o que no se cuenta con tanta información al respecto. Con todo la preocupación no es inexistente. Podemos destacar dos indicadores que nos parecen ilustrativos en este sentido. En primer lugar, la publicación de un trabajo pionero, reciente y muy ambicioso sobre la replicación en las Ciencias de la Educación (Makel & Plucker, 2014). En segundo lugar, el lanzamiento de una nueva revista de la AERA denominada *AERA Open*. En su editorial inaugural la revista se define como abierta a la inspección y la replicación y se hace eco de varios de los elementos asociados al debate de la transparencia y la reproducibilidad sobre los que volveremos más adelante (Warschauer, Duncan & Eccles, 2015).

Un marco general útil para entender el conjunto de preocupaciones que se vienen apuntando es el representado por la meta-investigación,

Recibido: 09-04-2017 Aceptado: 23-02-2018

disciplina o campo de trabajo emergente que se aproxima a la ciencia y sus prácticas "a vista de pájaro", esto es, adoptando una visión amplia y de conjunto (Ioannidis, Fanelli, Dunne & Goodman, 2015). La meta-investigación se preocupa de cómo se realiza, comunica, verifica, evalúa e incentiva la investigación científica. Y en cada una de esas áreas se focaliza en varios aspectos (ver tabla 1). No abordaremos aquí todos ellos, pero sí varios. Específicamente los vinculados con las prácticas de investigación cuestionables (en el área de métodos), la reproducibilidad y las nuevas prácticas en la publicación y difusión de la investigación científica (en el área de informes). Todo ello con especial atención a las Ciencias Sociales, del Comportamiento y de la Educación. Sobre evaluación y reconocimiento también haremos algunas reflexiones en el apartado de conclusiones[2].

TABLA I. Meta-investigación: áreas temáticas

Area	Intereses específicos ilustrativos
Métodos (realización)	Sesgos y prácticas cuestionables en la realización de investigación y métodos para su reducción; meta-análisis, síntesis de investigación, integración de evidencia; consorcios científicos y equipos colaborativos; integridad investigadora y ética.
Informes (comunicación)	Sesgos y prácticas cuestionables en la publicación y difusión de investigación; gestión de los conflictos de interés; registro de estudios y otros métodos para monitorizar y reducir sesgos y prácticas inadecuadas
Reproducibilidad (verificación)	Obstáculos a las prácticas de datos y métodos compartidos; estudios de replicación; replicabilidad y reproducibilidad de la investigación publicada y métodos para mejorarlas; efectividad de la corrección y auto-corrección de la literatura científica y métodos para mejorarla.
Evaluación (evaluación)	Efectividad, coste y beneficios de nuevas y viejas formas de revisión por pares y otros medios de evaluación de la ciencia y métodos para mejorarlos.
Incentivos (reconocimiento)	Precisión, efectividad, costes y beneficios de viejos y nuevos enfoques para evaluar y ordenar el rendimiento, la calidad y el valor de la investigación de individuos, grupos e instituciones.

Fuente: Adaptado de Ioannidis, Fanelli, Dunne & Goodman (2015)

(2) Una fuente de información útil sobre meta-investigación en general, con cierto énfasis en el área biosanitaria, es la página del Meta-Research Innovation Center at Stanford – METRICS (https://metrics.stanford.edu/). En el ámbito particular de las Ciencias Sociales puede verse: Berkeley Initiative for Transparency in the Social Sciences - BITSS (http://www.bitss.org/).

Recibido: 09-04-2017 Aceptado: 23-02-2018

Sobre sesgos y prácticas de investigación cuestionables

En los últimos años se ha ido generando un cuerpo creciente de literatura sobre sesgos y sobre errores que los investigadores cometen en los procesos de recogida, análisis, archivo y difusión de sus datos y resultados (para una visión de conjunto con énfasis en la investigación psicológica, véase p.e. Sijtsma, 2016 y los artículos de discusión asociados); de modo que en la actualidad se cuenta con una razonable evidencia de la alta prevalencia de algunas prácticas de investigación cuestionables (p.e. para el ámbito de las ciencias cognitivas véase Ioannidis et al., 2014).

En el gráfico I se muestra una versión ideal del modelo hipotético-deductivo del método científico. A cada etapa pueden asociarse amenazas potenciales que, tomadas conjuntamente, pueden contribuir a minar la robustez de la investigación publicada y pueden tener también impacto sobre la capacidad de la ciencia para la autocorrección. Algunos de los fenómenos que han recibido más atención por su alta prevalencia son (Munafò et al., 2017):

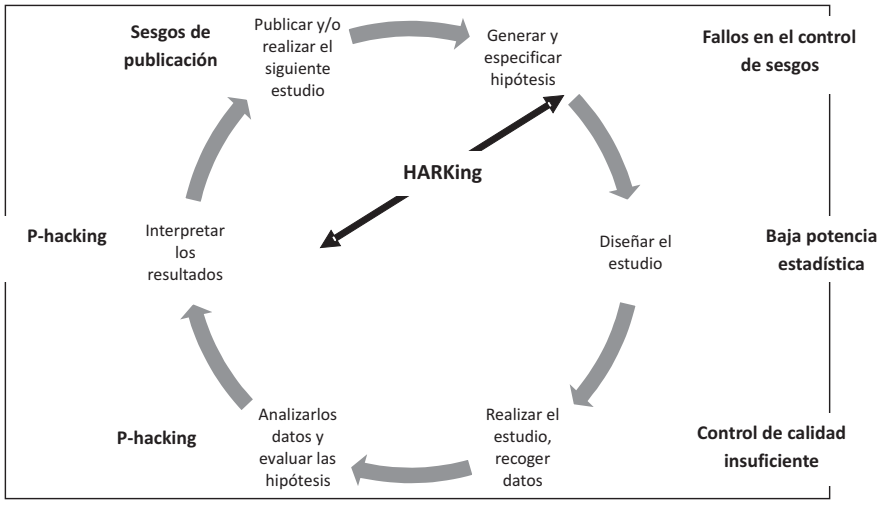
- *P-hacking,* también conocido como informe selectivo. Ha centrado gran parte del debate sobre la *crisis estadística de la ciencia*. Ocurre cuando los investigadores intentan varios análisis estadísticos o varias especificaciones sobre la elegibilidad de los datos buscando resultados estadísticamente significativos y, entonces, cuando los hallan en alguno de los múltiples escenarios que se han explorado, sólo se informa de éstos últimos. Las prácticas identificadas como más comunes en este grupo incluyen, entre otras:
  - Prescindir de condiciones experimentales que producen resultados inconsistentes con lo esperado y únicamente analizar e informar de los resultados del subconjunto de condiciones restantes; combinar o separar tratamientos a posteriori, una vez obtenidos los resultados; no informar de todas las variables analizadas, sólo de aquéllas asociadas a resultados deseados; reconstruir a posteriori una variable dependiente, por ejemplo usando únicamente algunos ítems, tras probar efectos significativos en esa re-definición operativa y no informar de ello.
  - Eliminar o no casos extremos u otras observaciones en función de las re-evaluaciones de la hipótesis nula; o incluir o excluir covariables a la vista de los resultados y no informar de ello.

Recibido: 09-04-2017 Aceptado: 23-02-2018

– Realizar análisis mientras se recogen los datos de un experimento o estudio y detenerse cuando se obtiene un resultado estadísticamente significativo, esto es, incrementar paso a paso el tamaño de la muestra, evaluando el estadístico de interés en cada paso, hasta que éste alcanza la significación y entonces detener la recogida de datos.

- *HARKing*, término empleado para designar la formulación a posteriori de las hipótesis de investigación: esto es, una vez obtenido el resultado. El escenario más común es informar de un estudio exploratorio como si fuera confirmatorio, en definitiva. Por ejemplo, no informar de las re-especificaciones realizadas sobre el modelo original en el modelo finalmente presentado.
- *Sesgos de publicación*. Los investigadores tienden a no considerar la publicación de trabajos donde no se obtuvieron resultados estadísticamente significativos, lo que está estrechamente vinculado con el ecosistema y la cultura de la publicación científica contemporánea.

GRÁFICO I. Amenazas a la reproducibilidad en el proceso de investigación científica



Fuente: Munafò et al. (2017)

Recibido: 09-04-2017 Aceptado: 23-02-2018

La cuestión, una vez establecida la existencia de tales amenazas para una acumulación fiable y robusta de la evidencia científica, es cómo combatirlas. Las soluciones posibles y las propuestas son múltiples, de naturaleza variada y situadas a distintos niveles en el esquema meta-investigativo que mostramos en la tabla I. En los epígrafes que siguen trataremos de tres aspectos, no exhaustivos pero sí muy incluyentes y muy presentes en la literatura como líneas de acción útiles.

Alternativas y "nuevos" énfasis en el análisis estadístico de datos científicos

En el inventario de prácticas de investigación cuestionables se incluye muy fundamentalmente prácticas que inflaccionan el error tipo I, esto es, que aumentan los falsos-positivos, por lo que su análisis está fuertemente asociado al debate de casi siete décadas sobre el contraste de la hipótesis nula (Balluerka, Gómez et al., 2005; Fernández-Cano & Fernández-Guerrero, 2009; Harlow, Mulaik & Steiger, 2016). Aunque creemos que se trata de un tema bien conocido, su mención parece imprescindible, puesto que en el contexto de la crisis de la reproducibilidad el debate ha resurgido con un vigor renovado.

Ha sido muy señalado que los ecosistemas de la producción y la publicación científica gravitan de un modo casi exclusivo sobre el modelo de investigación centrado en el contraste de hipótesis nula y por tanto generan fuertes incentivos para la obtención de resultados estadísticamente significativos y para la publicación exclusiva de los mismos, lo que tiene un vínculo directo con los sesgos de publicación y el informe selectivo. De hecho hay quien ha atribuido directamente la crisis de la reproducibilidad al uso de la pruebas de significación y defienden abierta y sencillamente su abandono. Un representante señalado de esta posición es Cummings, promotor de la llamada *nueva estadística* (Cumming, 2014), que promueve el uso exclusivo de la estimación frente al test de hipótesis y enfatiza tres elementos: el uso de tamaños de efecto, intervalos de confianza y meta-análisis. Las posiciones más extremas han tenido su impacto. El más llamativo, sin duda, lo representa el hecho de que por primera vez una revista de Psicología no marginal, indexada en JCR (*Basic and Applied Social Psychology*) en el año 2015 requiriera expresamente no usar contrastes de hipótesis nula en sus originales (Trafimow & Marks, 2015). También ha resurgido con

intensidad la propuesta de adoptar el enfoque bayesiano de la inferencia estadística (Mulder & Wagenmakers, 2016). Pero estas propuestas, digamos extremas, también han sido ampliamente contestadas desde posiciones más moderadas (y a nuestro juicio particular también más ponderadas) que defienden que el uso correcto de los test de hipótesis y su interpretación cuidadosa son una herramienta clave en el análisis de datos científicos (Morey, Rouder, Verhagen & Wagenmakers, 2014). De hecho se ha calificado de "pista falsa" la exclusiva atención prestada a este asunto en el contexto de la crisis, como si la sola eliminación del contraste de hipótesis nula resolviera todos los problemas y sus alternativas garantizaran por sí solas la validez, robustez y avance efectivo de la ciencia (Savalei & Dunn, 2015).

En este contexto cabe entender la publicación de una reciente declaración institucional de la American Statistical Association sobre el uso de la significación estadística y los p-valores en la investigación científica (Wasserstein & Lazar, 2015). Realmente, como se reconoce en el propio texto, no hay nada nuevo, pero es un buen indicador del estado del debate en el ámbito científico, que parece haber obligado a un pronunciamiento formal de este tipo por parte de la ASA.

En todo caso, como fruto de este renovado debate, se ha dibujado un espacio de amplio consenso sobre la necesidad de revisar algunas prácticas y adoptar definitivamente otras. Entre estas últimas destaca la necesidad de informar siempre del tamaño del efecto y de su precisión, mediante el correspondiente intervalo de confianza, práctica recomendada desde hace más de tres décadas. Y también la de consolidar un razonamiento netamente meta-analítico cuando se aborda la evidencia disponible sobre un problema, interpretando más cabalmente la información limitada que naturalmente puede derivarse de un estudio cualesquiera particular.

El asunto del tamaño del efecto ciertamente no es un tema nuevo. Las recomendaciones en este sentido se remontan a los años 80 (Thompson, 2008), fueron ampliamente difundidas por la APA a fines de los años 90 (Wilkinson & The Task Force on Statistical Inference, 1999), están presentes en todos sus manuales de publicación desde 2001 (p.e. APA, 2001; APA, 2010) y también en los estándares para la publicación de investigación empírica de la AERA (AERA, 2006). Pero realmente se está tratando como si lo fuera. Cabe preguntarse cuál ha sido el grado de adopción de las recomendaciones. Porque quizá ahí esté la explicación de por qué un *tema viejo* sigue siendo actual en algún sentido.

Recibido: 09-04-2017 Aceptado: 23-02-2018

Es interesante en el contexto norteamericano el trabajo de Peng, Chen, Chiang & Chiang (2013). Las autoras compilaron y analizaron 31 revisiones previas que se habían centrado en evaluar el uso e informe del tamaño del efecto en las revistas norteamericanas de Psicología y Educación. A esta minuciosa e ingente revisión de revisiones añaden su propia revisión de 451 artículos publicados en 2009 y 2010. El conjunto de 32 revisiones se dividen en dos periodos: antes y después de 1999 (fecha de publicación del conocido informe de Wilkinson y la TFSI). Se revisaron en total 116 revistas, incluyendo muchas de las asociadas a la APA y las asociadas a AERA.

De un modo apretado puede decirse que este trabajo pone de manifiesto dos cosas. En primer lugar, se identificó globalmente un incremento del uso de los índices del tamaño del efecto, entendido como un aumento en la tasa media de presencia en los artículos de investigación, que sigue siendo no obstante modesta. Aunque se registra una gran variedad entre revistas, las tasas globales pueden ser ilustrativas. Antes de 1999 la tasa media fue igual a 29,9% (mediana 29,4%) y después de 1999 igual a 54,7% (mediana 58%). En segundo lugar, se identificaron unas prácticas aún deficientes en su estimación y en su interpretación, entre las que destaca el escaso uso de los intervalos de confianza para los tamaños del efecto. Parece claro que la adopción de las recomendaciones resulta lenta y no fácil.

En el ámbito de la investigación educativa española no hemos podido encontrar trabajos publicados sobre el uso de medidas de tamaño del efecto y de sus intervalos de confianza. En Psicología sí hemos podido localizar dos trabajos que pueden ofrecer algún indicador siquiera grueso del estado de la cuestión.

García, Campos y De la Fuente (2011) revisaron 787 artículos que se publicaron en cuatro revistas españolas entre los años 2003 y 2008 (*Psicothema*, *Spanish Journal of Psychology*, *Psicológica, Internacional Journal of Clinical and Health Psychology*). Se encontró un porcentaje de reporte, respectivamente, del 21%, 32%, 3% y 19% en cada una de las revistas citadas. Como resultado final, y evaluando el total de publicaciones, se tuvo un porcentaje de aplicación del 21%. Caperos y Pardo (2013) analizaron los artículos publicados en 2011 en cuatro revistas españolas de Psicología indexadas en la base de datos JCR (*Anales de Psicología, Psicológica, Psicothema,* y *Spanish Journal of Psychology*). Sus resultados indican que el 41% de los artículos incluyó alguna medida del tamaño

del efecto. Conviene notar que estas dos revisiones no incluyeron el uso de los intervalos de confianza para el tamaño del efecto, recomendación presente ya en la edición de 2010 de las normas APA y que previsiblemente llevarán un tiempo complementario o adicional.

Parece claro que hay un amplio margen para la mejora. También que no parece plausible encontrar tasas mucho mayores en el ámbito de la investigación educativa.

La prevalencia de sesgos estadísticos, las malas prácticas y la lentitud registradas en la adopción de innovaciones estadísticas y prácticas mejoradas pueden sin duda ser usadas para desacreditar la validez de la ciencia. Sin embargo, creemos que el uso de métodos estadísticos rigurosos y su interpretación cuidadosa constituye una característica esencial, poderosa y sólida de la buena ciencia y una herramienta clave para sustentar su integridad. Y es en este sentido en el que se trata de avanzar.

Impulso renovado a la replicación y la reproducibilidad: el caso de las Ciencias de la Educación y del Comportamiento

En el ámbito de las Ciencias Sociales los trabajos de réplica han sido muy infrecuentes a pesar de su contribución clave al avance del conocimiento científico (Makel & Plucker, 2014). De hecho hasta hace unos pocos años no se contaba con una revisión sistemática de la replicación en investigación educativa.

En el trabajo de Makel & Plucker (2014) ya citado se analizaron las 100 revistas incluidas en la categoría *Education & Educational Research* del JCR con mayor factor de impacto de 5 años (edición 2011) y se analizaron todos los artículos que incluían el término de búsqueda *replicat** en el texto, considerando la historia completa de cada publicación.

La tasa general de replicación encontrada en el estudio fue igual a 0,13% (221/164.589 artículos), ocho veces menor que la estimada en Psicología por estos mismos autores (Makel, Plucker & Hegarty, 2012). También se identificó una evolución positiva de la misma en perspectiva temporal. Desde 1990 la tasa es casi cuatro veces mayor. O en otras palabras, los estudios de réplica han pasado de ser 1 de cada 2000 artículos de educación a casi 1 de cada 500.

Recibido: 09-04-2017 Aceptado: 23-02-2018

En lo que se refiere al análisis de los resultados, la tasa general de réplicas exitosas fue igual a 67,4%. Estos resultados, que indican que la mayoría de los estudios fueron exitosamente replicados, si bien difieren de los hallados en el ámbito de la Medicina y las Ciencias de la Salud (bastante menos halagüeños), parecen coincidir con los hallados en Psicología. No obstante los autores llaman la atención sobre el hecho de que casi la mitad de las réplicas (48,2%) fueron llevadas a cabo por los mismos equipos que publicaron el estudio original. Y la tasa de éxito en la réplica fue mayor cuando estuvo a cargo del mismo equipo o hubo cierto solapamiento en la autoría que cuando la llevó a cabo un equipo completamente independiente. Conviene entonces considerar las posibles limitaciones asociadas a las auto-réplicas, pues asiste la duda de si la mayor tasa de éxito es fruto del mayor conocimiento y experiencia o si se replican también los sesgos y las prácticas de investigación cuestionables informadas por la literatura en Ciencias Sociales. Esta cuestión requiere por tanto de investigación adicional.

En todo caso parece claro que la replicación es extremadamente infrecuente y que debe ser potenciada. Como señalan los autores de este trabajo, la replicación no es la panacea, no resolverá todos los retos y problemas sobre el rigor, la fiabilidad, la precisión y la validez de la investigación educativa. Sin embargo seguirla ignorando implícita o explícitamente indica una honda falta de comprensión sobre la ciencia y sobre cómo opera su avance.

Estas consideraciones han calado en el ámbito de la Psicología, donde la crisis como señalamos ha sido muy honda y donde se han desarrollado en los últimos años trabajos específicamente centrados en la replicación y la reproducibilidad. Comentaremos brevemente el trabajo ya citado de la *Open Science Collaboration,* publicado en *Science*. Este artículo tuvo y sigue teniendo un enorme impacto (1225 citas en Google Scholar el 1 de septiembre de 2017), ha originado un interesante debate académico con críticas y contra críticas y de algún modo ha venido a constituirse en un clásico en apenas dos años.

Se trata de un ambicioso proyecto colaborativo a gran escala, en la que participaron equipos de investigación independientes de todo el mundo, para obtener una estimación inicial de la reproducibilidad de la ciencia psicológica. Concretamente se llevaron a cabo réplicas de 100 estudios correlacionales y experimentales publicados en 2008 en tres revistas de Psicología de alto impacto (*Psychological Science*, *Social Psychology*,

Experimental Psychology: learning, memory and cognition) usando diseños con alta potencia y los materiales originales cuando estuvieron disponibles. Se usaron 5 indicadores para evaluar el éxito de la réplica: significación y valor-p, tamaños del efecto, evaluación subjetiva de los equipos de replicación y meta-análisis de los tamaños del efecto. Los resultados referidos a estos cinco indicadores son complejos y difíciles de resumir en unos párrafos, pues además incluyen análisis de posibles factores asociados a la reproductibilidad y se llevaron a cabo estudios exploratorios para inspirar la investigación futura. Pero algunos datos pueden ser bien ilustrativos. El 97% de los estudios originales obtuvieron resultados significativos frente al 36% de los correspondientes a las réplicas. El tamaño del efecto medio estimado en las réplicas fue igual a. 197, la mitad del correspondiente a los estudios originales (.403), lo que señala un descenso sustancial. Finalmente, el 39% de los efectos fueron valorados subjetivamente como replicados por los equipos de investigación.

Tomados en conjunto, estos y el resto de los resultados ofrecen una conclusión clara a juicio de los autores: una gran proporción de réplicas produjeron una evidencia más débil para los resultados originales, a pesar de usar los materiales proporcionados por los autores, de revisar anticipadamente la fidelidad metodológica y de la alta potencia estadística para detectar los tamaños del efecto originales.

En todo caso, como los propios autores apuntan, este trabajo deja muchas cuestiones abiertas, y debe verse como un primer paso, no como un punto de llegada. Los resultados originales ofrecieron evidencia tentativa, las réplicas añaden evidencia adicional, confirmatoria. En algunos casos las réplicas incrementan la confianza en los resultados originales. En otros casos más bien sugieren que se necesita más investigación para establecer la validez de los resultados originales. Y así es como cabalmente deberían ser entendidos estos esfuerzos.

Ciencia Abierta

El acuerdo unánime en la necesidad de aumentar la transparencia y la apertura en los procesos mediante los que se realiza, difunde y evalúa la investigación científica (como factores clave para garantizar la reproducibilidad) ha movido en los últimos años a diseñar y poner en

Recibido: 09-04-2017 Aceptado: 23-02-2018

marcha iniciativas variadas y que en conjunto vienen conformando lo que se conoce como Ciencia Abierta u *Open Science* (Morey et al., 2016; Nosek et al., 2015). U**na visión de conjunto y una buena ilustración del tipo de iniciativas que se van abriendo paso son las promovidas por el *Center for Open Science*, organización estadounidense sin ánimo de lucro orientada específicamente al desarrollo de herramientas tecnológicas y soportes gratuitos y de libre acceso para la promoción de la ciencia abierta y la reproducibilidad (**https://cos.io/**). Describimos a continuación tres de sus proyectos, especialmente representativos de los avances en este campo.**

Directrices para la promoción de la transparencia y la apertura (*Transparency and Opennes Promotion- TOP Guidelines*)

Esta propuesta incluye estándares para ocho módulos que pueden ser adoptados y desarrollados libremente en parte o totalmente por las revistas científicas. Cada módulo además incluye tres niveles de exigencia creciente en términos de transparencia. Se define así un marco flexible que reconoce que no todos los estándares son aplicables a todas las revistas o todas las disciplinas. Los módulos son:

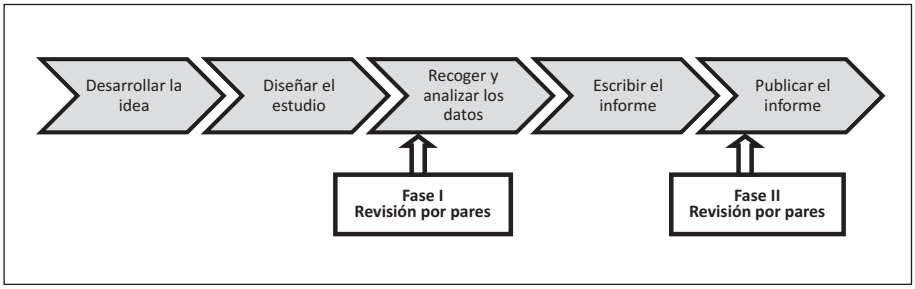
- Estándares de citación
- Transparencia de datos
- Transparencia de métodos analíticos (código)
- Transparencia de materiales de investigación
- Transparencia de análisis y de diseño
- Pre-registro de estudios
- Pre-registro de planes de análisis
- Replicación

Hay un número creciente de revistas adoptando este marco en campos disciplinares variados. Por ejemplo, el 15 de marzo de 2017 se adhirió la revista *Nature* y las revistas asociadas publicadas por Springer y Biomed Central.

Un tema de especial interés nos parece el de los artículos e informes pre-registrados. Constituyen un nuevo formato de publicación que enfatiza la importancia de la pregunta de investigación y la calidad de la

metodología mediante la revisión por pares previa a la recogida de datos (Nosek & Lakens, 2014). Concretamente se trata de aceptar protocolos de investigación de alta calidad provisionalmente, que serán publicados con independencia de los resultados obtenidos si los autores siguen la metodología especificada (ver gráfico II).

GRÁFICO II. Proceso de pre-registro de artículos y planes de análisis



Fuente: Adaptado de Center for Open Science (https://cos.io/rr/)

El formato está diseñado para incentivar las buenas prácticas y elimina una variedad de prácticas de investigación cuestionables, tales como la baja potencia estadística, el informe selectivo de resultados o los sesgos de publicación, mientras permite que se complete con flexibilidad el informe con resultados no previstos, calificándolos como tales. Se enfatiza así la distinción entre investigación exploratoria y confirmatoria para una adecuada valoración y evaluación de los resultados de la investigación científica.

En el momento de preparar este artículo más de medio centenar de revistas usan este formato, bien incluyéndolo como una forma regular en su política de aceptación de originales o bien como parte de números especiales.

Marco de Ciencia Abierta (*Open Science Framework - OSF*)

Entorno *web* abierto y gratuito que permite conectar y dar soporte al flujo de trabajo de la investigación a lo largo de todo el ciclo de un

Recibido: 09-04-2017 Aceptado: 23-02-2018

proyecto, pudiendo ser usado para colaborar, documentar, archivar, compartir y registrar proyectos de investigación, materiales y datos. Hay otras aplicaciones y plataformas, pero esta es una buena ilustración del tipo de herramientas y software innovador desarrollados para dar soporte a las nuevas prácticas de ciencia abierta. Algunas de las revistas que, por ejemplo, recomiendan este marco como repositorio para compartir datos y materiales y para el pre-registro de artículos son: *Biomed Central journals, Cognition, Nature, Psychological Science, Perspectives on Psychological Science, PloS, Science, Scientific Data*.

Distintivos de Ciencia Abierta (*Open Science Badges*)

Se usan para identificar y reconocer expresamente a los artículos que cumplen con determinados estándares de transparencia y apertura. Pueden otorgarse sobre la base de la simple declaración del autor o como resultado de un proceso de revisión por pares, de acuerdo con la política editorial que fije cada revista.

Cuando preparamos este artículo una veintena de revistas habían adoptado estos distintivos. A título de ejemplo, los requisitos que se exigen para otorgar el distintivo de datos abiertos son:

- Proporcionar una URL, DOI, u otro *path* permanente para acceder a un repositorio de acceso público y abierto.
- Proporcionar junto con los datos el conjunto de información suficiente para que un investigador independiente pueda reproducir los resultados informados en el artículo (metadatos).
- Contar con un tipo de licencia abierta que permita a otros copiar, distribuir y hacer uso de los datos aunque igualmente se permita mantener los derechos y el crédito a los propietarios cuando corresponda (*Creative Commons* ha definido recientemente varias licencias con este objetivo).

Con fecha 1 de agosto de 2017 la American Psychological Association – APA anunció un acuerdo con el *Center for Open Science* mediante el cual sus revistas ofrecerán distintivos de ciencia abierta y harán uso del Marco de Ciencia Abierta (OSF) como repositorio para sus datos y para la gestión de servicios de *preprint* (APA, 2017).

Conclusiones: algunos retos de futuro para la investigación educativa española

De la revisión realizada en los epígrafes precedentes cabe derivar algunos retos y desafíos que quizá deba enfrentar la investigación educativa en nuestro país. Vamos a ordenar algunos de ellos sobre cuatro dimensiones.

Análisis de las prácticas actuales

Al preparar este trabajo hemos podido identificar una cierta falta de tradición en la revisión y evaluación de las prácticas de investigación educativa en España, a diferencia de lo que ocurre en otros contextos, o en el nuestro, en otros ámbitos afines. En el ámbito de la Psicología pueden encontrar trabajos que, por ejemplo, analizan el uso de métodos estadísticos para identificar malas y buenas prácticas y elaborar recomendaciones de uso (véase, además de los trabajos ya citados de Caperos & Pardo, 2013 y García, Campos & de la Fuente (2011); el de Izquierdo, Olea y Abad, 2014, entre otros). Por tanto creemos que quizá tenemos aquí una primera cuestión pendiente. Nos falta información o evidencia que permita diagnosticar y en su caso tratar de corregir. El desarrollo de trabajos centrados en analizar las prácticas estadístico-metodológicas actuales en la investigación educativa española constituye entonces una línea de trabajo necesaria y oportuna.

Formación, información y sensibilización de los investigadores

Dada la alta prevalencia con la que se han encontrado prácticas de investigación cuestionables parece razonable preguntarse si realmente en todos los casos los investigadores somos realmente conscientes de incurrir en ellas. Considerar este escenario puede ayudar a entender mejor el asunto. Quizá entre los investigadores se da una infravaloración o falta de conciencia de la dificultad de los temas estadísticos, dificultades para entender, por ejemplo, el impacto sobre las tasas de error tipo I de las prácticas de investigación cuestionables. Centrarnos en la intencionalidad para distinguir la mala praxis (por falta de información o formación) del

Recibido: 09-04-2017 Aceptado: 23-02-2018

fraude, como ha sido señalado en la literatura (Sijtsma, 2016), puede sugerir la necesidad de intervenir sobre los procesos de formación inicial y permanente de los investigadores en el ámbito estadístico-metodológico. Parece entonces pertinente una cierta reflexión sobre la formación de postgrado inicial y también sobre la oportunidad de arbitrar cauces de formación permanente para los investigadores en activo. Porque las dificultades y la lentitud en la adopción de innovaciones y buenas prácticas renovadas por parte de los investigadores es un asunto que ha recibido una atención creciente (Henson, Hull & Williams, 2010; Sharpe, 2013; Cohen, 2017) y creemos que quizá requiera de consideración también en nuestro contexto.

Políticas editoriales: el papel de editores y revisores

En todo caso los investigadores queremos publicar y por tanto puede pensarse que estamos bien dispuestos a responder a los requisitos que ello exige. De ahí que en principio las políticas editoriales de las revistas científicas tengan un papel esencial en el modo en que se realiza y se difunde la investigación. Es razonable pensar que si las revistas hacen explícitos unos adecuados estándares estadístico-metodológicos para la aceptación de originales, asumidos y tomados como referente común por los revisores, no penalizan o invitan a la publicación de réplicas y fomentan o instauran políticas de ciencia abierta, ello tendrá un efecto positivo directo, como parece estar poniéndose de manifiesto en otros contextos (véase Munafò et al., 2017). En la práctica creemos que ello supone un reto fabuloso en nuestro entorno y en nuestro ámbito. En esta década, en nuestro país, hemos asistido a una consolidación extraordinaria de un buen número de revistas científicas de calidad en educación, con grandes y frecuentemente desinteresados esfuerzos por parte de la comunidad investigadora y muy especialmente de editores y revisores (Ruiz-Corbella, Galán & Diestro, 2014). Un buen número de publicaciones han ingresado en índices internaciones y sólo mantener los logros habidos creemos personalmente que ya es en sí un gran reto. No obstante, creemos que las líneas apuntadas señalan un camino de futuro posiblemente necesario que ya han emprendido importantes revistas de educación (véase p.e. Lopez, Valenzuela, Nussbaum & Tsai, 2015).

Políticas académicas y científicas: el papel de los financiadores y evaluadores de la producción investigadora

Por último, pero sin duda no menos importante, creemos que es esencial considerar el papel clave que tienen los agentes que financian y evalúan la producción científica de investigadores, equipos e instituciones, porque a la postre son los que diseñan el sistema de incentivos y reconocimiento que modula en gran medida las conductas individuales y colectivas. Si se prima en la evaluación de la producción exclusivamente la cantidad y la presión por publicar sigue caracterizando extraordinariamente el ecosistema en el que se desarrolla el trabajo de los investigadores, parece difícil que se puedan mejorar algunas de las prácticas citadas. Si publicar datos en abierto no es incentivado de ningún modo, posiblemente no será una práctica que se extienda fácilmente; o si la participación en proyectos de amplio alcance colaborativos y los consorcios de investigación no se incentivan, resultará difícil que se introduzcan y consoliden en la práctica. Coincidimos por tanto con aquéllos que han señalado que los cambios en este nivel institucional son clave (Smaldino & McElreath, 2016) y es donde de hecho se concentran gran parte de los retos y desafíos que se han venido apuntando a lo largo de este artículo.

Reflexión final

En esencia lo que hemos querido subrayar en este trabajo es que el progreso científico es un proceso acumulativo de reducción de la incertidumbre (más que de la obtención de certezas) y que este proceso sólo puede tener éxito si la ciencia misma opera sobre la base de un natural y sistemático escepticismo sobre sus propios hallazgos. Creemos que el análisis hecho pone de manifiesto que hay un claro margen para la mejora de las prácticas científicas. Pero también esperamos haber puesto suficientemente de manifiesto que ante las hipótesis, más que plausibles, de que el ecosistema de la ciencia y su cultura actual pueden estar afectando negativamente a la validez y reproducibilidad de sus resultados, el sistema científico se está comportando como cabía esperar: revisando sistemática y críticamente sus prácticas, formulando alternativas de mejora y diseñando nuevos mecanismos para la auto-corrección. Una

Recibido: 09-04-2017 Aceptado: 23-02-2018

mirada como la que hemos propuesto aquí creemos que puede ser útil para contribuir a la reflexión sobre nuestras propias prácticas y al avance, en definitiva, de la investigación educativa en nuestro país.

Referencias bibliográficas

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. doi:10.3102/0013189X035006033.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2017, August 1). *APA Journals Program collaborates with Center for Open Science to advance open science practices in psychological research*. Retrieved from: http://www.apa.org/news/press/releases/2017/08/open-science.aspx.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452-454. doi:10.1038/533452a

Balluerka, N., J. Gómez, et al. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 1(2): 55-70.

Caperos, J.M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414.doi: 10.7334/psicothema2012.207.

Cohen, B. H. (2017). Why the resistance to statistical innovations? A comment on Sharpe (2013). *Psychological Methods, 22*(1), 204-210. doi: 10.1037/met0000058.

Cumming, G. (2014).The new statistics: why and how. *Psychological Science, 25*,7–29.doi:10.1177/0956797613504966.

Fernández-Cano, A., & Fernández-Guerrero, I. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*. Madrid: La Muralla.

García, J. Campos, E., & De la Fuente, L. (2011). The use of the effect size in JCR Spanish journals of Psychology: from theory to fact .*The Spanish Journal of Psychology*, *14*(2), 1050-1055.

Goodman, S.N. (2016). Aligning statistical and scientific reasoning. Misunderstanding and misuse of statistical significance impede science. *Science, 352* (6290), 1180-1181. doi: 10.1126/science.aaf5406.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (2016). *What if there were no significance tests?* Classic Edition. New York: Routledge.

Henson, R.K., Hull, D.M., & Williams, C.S. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educational Researcher, 39*(3), 229-240. doi: 10.3102/0013189X10365102.

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLoS Biology*, *13*(10),1-7. doi: 10.1371/journal.pbio.1002264.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences, 18*(5), 235-241.

Izquierdo, I., Olea, J., and Abad, F.J. (2014). Exploratory Factor Analysis in validation studies: uses and recommendations. *Psicothema, 26*(3), 395-400.

Ledgerwood, A. (2014). Introduction to the Special Section on Advancing Our Methods and Practices *Perspectives on Psychological Science,* 9(3) 275–277. doi: 10.1177/1745691614529448.

Lopez,X., Valenzuela, J., Nussbaum, M. & Tsai, C.(2015). Some recommendations for the reporting of quantitative studies (Editorial). *Computers & Education, 91*,106-110.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304 –316. doi: 10.3102/0013189X14545513.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542. doi:10.1177/1745691612460688.

McNutt, M. (2014). Journals unite for reproducibility (Editorial). *Science, 346*(6210), 679-679. doi:10.1038/515007a.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming (2014). *Psychological science*, *25*(6), 1289-1290.

Recibido: 09-04-2017 Aceptado: 23-02-2018

Morey, R.D. et al. (2016). The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science, 3*, 150547. doi: http://dx.doi.org/10.1098/rsos.150547.

Mulder, J. & Wagenmakers, E.J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology, 72*, 1–5.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. doi:10.1038/s41562-016-0021.

Nature (2014). Journals unite for reproducibility (Editorial). *Nature, 515*(7525), 7.

Nosek, B. A., et al. (2015). Promoting an open research culture. *Science, 348*, 1422-1425. DOI: 10.1126/science.aab2374.

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

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

Peng, C-Y., Chen, L-T., Chiang, H-M., & Chiang, Y-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*, 157-209. doi: 10.1007/s10648-013-9218-2.

Peng, R.D. (2015). The reproducibility crisis in science. A statistical counterattack. *Significance*, 30-32

Perspectives on Psychological Science. (2012). Special section on replicability in psychological science: A crisis of confidence? Retrieved from: http://pps.sagepub.com/content/7/6.toc

Perspectives on Psychological Science. (2014). Special section on Advancing our methods and practices. Retrieved from: http://journals.sagepub.com/toc/ppsa/9/3.

Ruiz-Corbella, M., Galán, A. & Diestro, A. (2014). Las revistas científicas de Educación en España: evolución y perspectivas de futuro. *RELIEVE, 20* (2), art. M1. doi: 10.7203/relieve.20.2.436.

Savalei V., Dunn E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology, 6*, 245. doi: 10.3389/fpsyg.2015.00245.

Recibido: 09-04-2017 Aceptado: 23-02-2018

Schweinsberg, M. et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55-67. doi: https://doi.org/10.1016/j.jesp.2015.10.001.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*(4), 572. doi: 10.1037/a0034177.

Sijtsma, K. (2016). Playing with data–Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*(1), *1-15*. doi:10.1007/s11336-015-9446-0.

Sijtsma, K., Veldkamp, C.L.S. & Wicherts, J.M. (2016). Improving the conduct and reporting of statistical analysis in Psychology. *Psychometrika, 81*(1), 33-38. doi:10.1007/s11336-015-9444-2.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi:10.1177/0956797611417632.

Smaldino, P.E., McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 160384. doi:http://dx.doi.org/10.1098/rsos.160384.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes (pp. 246-262). En Osborne, J.W. (2008)(Ed.). *Best Practice for Quantitative Methods*. London: Sage.

Trafimow, D. & Marks (2015). Editorial. *Basic and Applied Social Psychology, 3*7 (1), 1–2.

Waldman, I. D., & Lilienfeld, S. O. (2015). Thinking about data, research methods, and statistical analyses: Commentary on Sijtsma's (2014) "Playing with data". *Psychometrika*, *81*(1), 16-26. doi:10.1007/s11336-015-9447-z.

Warschauer, M., Duncan, G. J., & Eccles, J. S. (2015). Inaugural Editorial: what we mean by "open". *AERA Open,1* (1),1-2. doi: 10.1177/2332858415574841.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician, 70*(2), 129-133. doi: 10.1080/00031305.2016.1154108.

Wigboldus, D. H. J., & Dotch, R. (2015). Encourage playing with data and discourage questionable reporting practices. *Psychometrika, 81*(1), 27-32. doi:10.1007/s11336-015-9445-1.

Recibido: 09-04-2017 Aceptado: 23-02-2018

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594.

Información de contacto: Ángeles Blanco Blanco. Universidad Complutense de Madrid, Facultad de Educación, Departamento de Investigación y Psicología en Educación. Rector Royo Villanova s/n 28040 Madrid. E-mail: ablancob@ucm.es

State of scientific practices and educational research. Potential challenges for the next decade

Estado de las prácticas científicas e investigación educativa. Posibles retos para la próxima década

DOI: 10.4438/1988-592X-RE-2017-381-386

Ángeles Blanco-Blanco
Universidad Complutense de Madrid

Abstract

This paper presents a review of the state of current scientific practices and their potential impact on the quality of educational research. From a post positivist conception of scientific research in education, the matter is addressed in the general context of the current debate about science, its reliability, robustness and reproducibility. Theoretically and conceptually, the study adopts a meta-research approach. From a methodological perspective, a review of the literature on the subject is carried out that allows a reasoned reflection on the *status quo* of scientific practices. This is done by considering some of the most relevant papers published in recent years on meta-science in general, and in the field of Education and Behavioral Sciences. In the first place, the so-called *crisis of science* is characterized, and particularly the presence of biases and questionable research practices in scientific research. Below are some of the key corrective elements proposed to strengthen and enable a more effective advancement of the scientific enterprise. These include: alternative and "new" emphasis on the statistical analysis of scientific data; renewed impetus to replication and reproducibility; and new modes of production, dissemination and assessment of research associated with open science. The article closes with some reflections regarding possible challenges for Spanish educational research in the next decade. The conclusions are organized around four axes: the development of meta-research studies; training, information and awareness-raising of researchers about questionable research practices; updating editorial policies; and the role of funders and evaluators of scientific production.

Received: 09-04-2017 Accepted: 23-02-2018

Key words: Questionable research practices, Reproducibility crisis, Open Science, Scientific ecosystem, Meta-research, Educational research.

Resumen

El trabajo presenta una revisión del estado de las prácticas científicas actuales y de su impacto potencial en la calidad de la investigación educativa. Desde una concepción post-positivista de la investigación científica en educación, el problema se desarrolla en el contexto general del debate actual sobre la ciencia, sobre su fiabilidad, su robustez y su reproducibilidad. Teórica y conceptualmente el estudio adopta el enfoque propio de la meta-investigación. Desde el punto de vista metodológico, se lleva a cabo una revisión de la literatura sobre el tema que permita una reflexión fundamentada sobre el *status quo* de las prácticas científicas. Para ello se consideran algunos de los trabajos más relevantes y de mayor impacto publicados en los últimos años sobre meta-ciencia, en general y en el ámbito de las Ciencias de la Educación y del Comportamiento en particular. En primer lugar se caracteriza lo que se ha etiquetado como *crisis de la ciencia* y particularmente se trata de la presencia de sesgos y prácticas de investigación cuestionables en la investigación científica. A continuación se presentan algunos de los elementos correctores clave que se han propuesto para fortalecer y permitir un avance más efectivo de la empresa científica. Estos incluyen: alternativas y "nuevos" énfasis en el análisis estadístico de datos científicos; el impulso renovado a la replicación y la reproducibilidad; y los nuevos modos de producción, difusión y evaluación de la investigación ligados a la ciencia abierta. El artículo se cierra con algunas reflexiones relativas a posibles retos para la investigación educativa española en la próxima década. Las conclusiones se organizan en torno a cuatro ejes: el desarrollo de estudios meta-investigativos; la formación, información y sensibilización de los investigadores sobre prácticas de investigación cuestionables; la actualización de las políticas editoriales; y el papel de los financiadores y evaluadores de la producción científica.

Palabras clave: Prácticas de investigación cuestionables, Crisis de reproducibilidad, Ciencia abierta, Ecosistemas de la producción y la publicación científicas, Meta-investigación, Investigación educativa.

Empirical research in education is required to inform about educational practices and interventions. However, in order to be able to effectively contribute to the decision-making, one necessary condition, although not the only one, is that the educational research be scientifically robust, reliable and valid. This study aims to provide a systematic reflection on this condition. We present a review of current scientific practices

Received: 09-04-2017 Accepted: 23-02-2018

to study their potential impact on the quality and rigor of scientific research. The analysis is carried out from a post-positivist conception of scientific research in education, and in the general context of the current debate about science, its reliability, robustness and reproducibility. From a theoretical-conceptual perspective the study adopts a meta-research approach. From a methodological perspective, a literature review was conducted on the study area which permitted a well-founded reflection on the *status quo* of current scientific practices. This was carried out by considering some of the most relevant and highest impact studies published in recent years on meta-science in general, and more specifically on Educational and Behavioral Sciences.

Science in crisis and meta-research

In recent years, a wide range of disciplinary fields have been experiencing a crisis of confidence in what could be referred to as the *status quo* of scientific practices. There has been talk of a *crisis of science* and, especially, a *crisis of reproducibility* (Munafò et al., 2017). This situation is reflected by articles simultaneously published in the journals *Science* and *Nature* in 2014 under the revealing title: *Journals unite for reproducibility* (Nature, 2014; McNutt, 2014). The publishers describe a joint agreement to adopt new principles and guidelines for the publication of clinical research. However, fundamentally, and the point we would like to emphasize here, is that their ultimate goal was to encourage recent initiatives to respond to this general lack of trust arising from evidence of poor reproducibility, fraud and malpractice, which has been undermining the trust of experts, and the general public alike, in science.

The scientific community's concern about this is illustrated by the results of a recent survey published in *Nature* (Baker, 2016). Of the 1576 researchers completing the questionnaire, 90% considered a reproducibility crisis to exist and 52% regarded it to be significant. The causes most frequently mentioned included questionable research practices (e.g. selective reporting or *p-hacking*), the pressure to publish and matters related to use of the statistical system applied and the design of the research. This last statistical-methodological dimension heads the list of solutions mentioned by the survey respondents, especially in relation to improving training in, and supervision of, statistics and

Received: 09-04-2017 Accepted: 23-02-2018

methodology. In fact, it was mentioned that the *crisis* could be largely due to an incorrect use of statistics in scientific research (Goodman, 2016; Peng, 2015).

The difficulties encountered to reproduce the published work should first be regarded as a symptom, as the result of a complex problem linked to several factors associated with the setting and the parameters with which the science is generated and disseminated. As Makel and Plucker remind us (2014), the fact that a result is not reproducible does not make it false. Moreover, a reproducible result is not necessarily a correct one. However, reproducibility is a cornerstone of the scientific system and the debate surrounding it seems to have become focused, and rightly so, on the need for rigor and transparency in scientific practices[1].

This crisis has possibly been the most turbulent in recent years in the area of Psychology (Ledgerwood, 2014), an area that shares similarities with that of Education. The detection of fraudulent work conducted by respected researchers, the publication of dubious and controversial discoveries and evidence of questionable practices have all fueled the suspicion that perhaps all is not well behind the high success rates of researchers' predictions for results from their own studies. Hence, Simmons, Nelson & Simonsohn (2011) suggested that the hyper-precision in the prediction (e.g. rejection of the null hypothesis) is probably due to several factors including: collecting data only until the desired results have been reached, not reporting information about failed experiments, omission *a posteriori* of observations or variables that do not support the proposed hypothesis. In other words, a long list of questionable research practices.

All this has given rise to an extensive and lively debate in the area of psychology research in recent years. Although we do not give a detailed description here, this is well-illustrated by at least two indicators:

- The publication of numerous special sections on questionable research practices, replicability and reproducibility in renowned

(1) Indeed replication, understood as the repetition of previous studies by independent researchers, is the essential approach to collect scientific evidence that supports or refutes a study hypothesis. Replicability, therefore, is associated with the probability that an independent study produces results consistent with the original study. The concept of reproducibility is a recent variation in this concept (Peng, 2015). It refers to a minimum standard relating to research that entails the massive computation of data. It is defined as the possibility of recalculating the original results given the database and the original analytical specifications (e.g. code and documentation).

Received: 09-04-2017 Accepted: 23-02-2018

journals such as *Perspectives on Psychological Science* (2012, 2014) or *Psykometrica* (Sijtsma, 2016; Sijtsma, Veldkamp & Wicherts, 2016; Waldman & Lilienfeld, 2015; Wigboldus & Dotch, 2015).

- The encouragement of ambitious large-scale research programs specifically aimed at replicating the results of published psychology research, mainly adopting new collaborative work approaches between independent research teams and consortia (see Open Science Collaboration, 2015 or Schweinsberg et al., 2016).

In our opinion, this crisis is less evident in educational research, which does not mean that similar problems are not encountered. This could be due to a lack of scientific evidence of these occurrences or the existence of fewer data to support them, but the concern is still there. This is illustrated by two main indicators: First, the publication of a recent pioneering and highly ambitious work on replication in Education Sciences (Makel & Plucker, 2014); Secondly, the launch of a new journal by AERA called *AERA Open*. In the editorial of its first issue, the journal defines itself as being open to inspection and replication, echoing some of the elements associated with the debate on transparency and reproducibility that we will return to later (Warschauer, Duncan & Eccles, 2015).

A useful framework within which to contemplate all these problems is that of meta-research, an emerging discipline or study area that approaches science and its practices from a bird's eye view, in other words, by adopting a broader and more comprehensive perspective (Ioannidis, Fanelli, Dunne & Goodman, 2015). Meta-research focuses on how scientific research is conducted, communicated, verified, evaluated and promoted, and focuses on different aspects of each of these areas (see table 1).

Here, we will examine some of these, especially those related to questionable research practices (in the area of methods), reproducibility and new practices in the publication and dissemination of scientific research (in the area of reports), mainly focusing on Social, Behavior and Education sciences. We will also make some observations about evaluation and recognition processes in the conclusions section.

Received: 09-04-2017 Accepted: 23-02-2018

TABLE I. Meta-research: subject areas

Area	Specific examples of interest
Methods (implementation)	Biases and questionable practices in implementation of the research, methods to reduce such biases, meta-analysis, synthesis of the research, integration of evidence, scientific consortia and collaborative research teams, research integrity and ethics.
Reports (communication)	Biases and questionable practices in the publication and dissemination of research, management of conflicts of interest, recording of studies and other monitoring methods to reduce bias and inadequate practices.
Reproducibility (verification)	Obstacles to practices of shared data and methods, replication studies; replicability and reproducibility of the published research and methods to improve them, effectiveness of the correction and self-correction of scientific literature and techniques to improve them.
Evaluation	Effectiveness, cost and benefits of new and old methods of peer reviewing and other scientific assessment approaches and methods to improve them.
Incentives (recognition)	Precision, efficacy, costs and benefits of old and new approaches to evaluate and order the performance, quality and value of the research of individuals, teams and institutions.

Source: Adapted from Ioannidis, Fanelli, Dunne & Goodman (2015)

Biases and questionable research practices

The past few years have witnessed a growing body of literature about the biases and errors that researchers make during the collection, analysis, storage and dissemination of data and results (for a general overview with emphasis on psychology research see Sijtsma, 2016 and related articles); as a result, there is currently a reasonable amount of evidence for the high prevalence of some questionable research practices (e.g. for the area of cognitive sciences see Ioannidis et al., 2014).

Graph I shows an ideal version of the hypothetical-deductive model of the scientific method. Each stage can be associated with potential risks, which, taken together can help to undermine the robustness of the published research and can affect the capacity for self-correction of the science. Some of the phenomena receiving the most attention owing to their high prevalence include (Munafò et al., 2017):

- *P-hacking,* also known as selective reporting. This questionable practice has been central to the debate on the *statistical crisis of*

Received: 09-04-2017 Accepted: 23-02-2018

the science. This occurs when researchers try out different statistical analyses or different specifications for the eligibility of data in an attempt to find statistically significant results. When they obtain these in some of the numerous settings they have explored, only the successful ones are reported. Some of the most frequent practices identified in this group include:

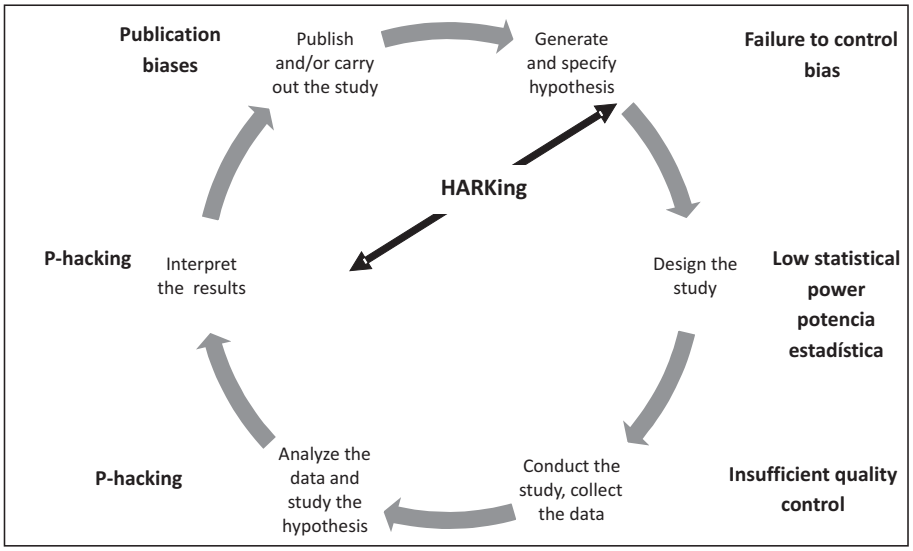
- Omitting experimental conditions that produce data inconsistent with expected results and only analyzing and reporting results of the subgroup of remaining conditions; combining or separating treatments *a posteriori*, after obtaining the results; not reporting all the variables studied but only those associated with the desired results; defining a dependent variable *a posteriori*, for example, only using some items after proving significant effects in this redefined operative measure, and not reporting this.
- Omitting or not omitting outliers or other observations in relation to re-assessments of the null hypothesis; or including or excluding covariables in view of the results obtained and not reporting this.
- To perform the analysis while collecting experimental or study data and to stop compiling data when a statistically significant result is obtained. In other words, to increase the size of a sample stepwise, assessing the statistical interest at each step, until reaching statistical significance and then not collecting any more data.

- *HARKing*, a term that refers to formulation of the research hypothesis *a posteriori*: after obtaining the result. The most common case is to report an exploratory study as being confirmatory. For example, to not report the respecifications of the original model that are later included in the final model.
- *Publication biases*. The researchers tend to not publish studies in which they have not obtained statistically significant results, this is closely linked to the environment and culture of contemporary scientific publication.

The question, after having established the existence of threats to a reliable and robust accumulation of scientific evidence, is how to overcome these. There are multiple possible solutions that have been proposed, varied in nature and situated at different levels in the meta-research diagram shown in Table I. In the following sections we focus

on three of these; although they do not cover the whole range they are inclusive and are very present in the literature as useful lines of action.

GRAPH I. Threats to reproducibility in the scientific research process.



Source: Munafò et al. (2017)

Alternative and "new" emphasis on the statistical analysis of scientific data

The list of questionable research practices mainly includes actions that inflate the type I error, in other words, that increase the false-positives, so their analysis is strongly associated with the almost seventy year controversy about null hypothesis testing (Balluerka, Gómez et al., 2005; Fernández-Cano & Fernández-Guerrero, 2009; Harlow, Mulaik & Steiger, 2016). Although this can be considered as a highly familiar area, this debate has returned with renewed vigor within the context of the reproducibility crisis.

Received: 09-04-2017 Accepted: 23-02-2018

It has been frequently reported that the ecosystems of scientific production and publication gravitate almost exclusively toward a research model based on null hypothesis testing. This creates a strong incentive to obtain statistically significant results, which are the only results published, which in turn directly link to publication bias and selective reporting. In fact, some authors have even directly blamed the reproducibility crisis on the use of significance tests, and openly and simply defend the idea of not using them. One noteworthy representative of this position is Cummings, an advocate of the so-called *new statistics* (Cumming, 2014), who promotes the exclusive use of estimation instead of hypothesis testing, identifying three key elements: the use of effect sizes, confidence intervals and meta-analysis. The most extreme positions have had some impact. The most noteworthy example is that of a reasonably mainstream psychology journal, indexed in JCR (*Basic and Applied Social Psychology*), which in 2015 expressly demanded that null hypotheses tests not be used in its original articles (Trafimow & Marks, 2015). There has also been an evident resurfacing of the Bayesian approach to statistical inference (Mulder & Wagenmakers, 2016). But these somewhat more extreme approaches have also been widely contested from more moderate, and in our opinion more carefully considered, standpoints, which suggest that the correct use of hypothesis testing and its careful interpretation are crucial to scientific data analysis (Morey, Rouder, Verhagen & Wagenmakers, 2014). In fact, exclusively focusing on this aspect has been described as a red herring in the context of the crisis, as if doing away with the null hypothesis test would solve all the problems, and that using alternative approaches instead would guarantee the validity, robustness and effective advancement of the science (Savalei & Dunn, 2015).

In this context, one can understand the publication of a recent institutional statement made by the American Statistical Association about the use of statistical significance and p-values in scientific research (Wasserstein & Lazar, 2015). In fact, as the text itself implies, there is nothing new, but the need for a formal announcement of this sort by the ASA clearly reflects the current state of the debate in the scientific setting.

In any case, this renewed debate has resulted in a broad consensus about the need to review some practices and adopt others. The latter include the need to always report the effect size and its precision, by recording the corresponding confidence interval, which has been the recommended practice for more than three decades. Also, consolidation

of a net meta-analytical reasoning when considering the evidence available for a given problem, recognizing the limited information than can be obtained from any single study.

The matter of effect size is certainly not a new one. Recommendations about this go back to the Eighties (Thompson, 2008), were widely disseminated in the 90's (Wilkinson & The Task Force on Statistical Inference, 1999), have been present in all their publication manuals since 2001 (e.g. APA, 2001; APA, 2010) and also in the standards for the publication of empirical research of the AERA (AERA, 2006). But it is indeed being treated as if it were new. It would be interesting to find out the extent to which these recommendations have been adopted. Perhaps this explains why an *old matter* is still to some extent also a current one.

The work of Peng, Chen, Chiang & Chiang (2013) is interesting in the North American context. The authors compile and analyze 31 previous reviews that had focused on assessing the use and reporting the effect size in North American journals of Psychology and Education. To this detailed and enormous review of reviews they add their own review of 451 articles published in 2009 and 2010. The group of 32 reviews is separated into two periods: before and after 1999 (the time of publication of the renowned Wilkinson report and the TFSI). A total of 116 journals were reviewed, including many of those associated with APA and AERA.

In summary, this work can be said to show two things. Firstly, an overall increase in the use of effect size indices was identified, understood as an increase in the mean rate of their presence in research articles, although this is still only modest. In spite of a great variability among the journals, the overall rates can give a general idea. Before 1999, the mean rate was around 29.9% (median 29.4%) and after 1999 it was around 54.7% (median 58%). Secondly, incorrect practices in their estimation and interpretation were reported, which included a limited use of confidence intervals for effect sizes. It is evident therefore that adoption of the recommendations is slow and difficult.

In the area of Spanish educational research we did not find any published studies on the use of effect size measurements and their confidence intervals. In Psychology, however, there were two rough indicators of the state of play.

García, Campos and De la Fuente (2011) published reviews of 787 articles in four Spanish journals between 2003 and 2008 (*Psicothema*, *Spanish Journal of Psychology*, *Psicológica, Internacional Journal of*

Received: 09-04-2017 Accepted: 23-02-2018

Clinical and Health Psychology). They found a percentage of reporting of 21%, 32%, 3% and 19%, respectively, in each of the cited journals. As a final result, taking into account all the publications, they obtained a percentage of application of 21%. Caperos and Pardo (2013) analyzed the articles published in 2011 in four Spanish journals of Psychology indexed in the JCR database (*Anales de Psicología, Psicológica, Psicothema,* and *Spanish Journal of Psychology*). Their results show that 41% of the articles included some measure of effect size. It is noteworthy that these two reviews did not include the use of confidence intervals for effect size, a recommendation which was already present in the 2010 issue of the APA regulations and whose adoption will probably take a complementary or additional time. There clearly appears to be much room for improvement. Moreover, we would not expect to find rates much higher than this in educational research.

The prevalence of statistical bias, malpractice and the very slow adoption of statistical innovations and improved practices can undoubtedly be used to discredit the validity of science. However, we believe that progress in the use of rigorous statistical methods and their careful interpretation are essential, powerful and solid characteristics of good science and key to upholding its integrity.

A renewed boost to replication and reproducibility: the case of sciences of education and behavior

In the area of Social Sciences replication studies are uncommon in spite of the key contribution they make to the advancement of scientific knowledge (Makel & Plucker, 2014). In fact, until just a few years ago there had been no systematic review of replication in educational research.

In the work of Makel & Plucker (2014), cited previously, for the 100 journals included in the category *Education & Educational Research* of the JCR with an impact factor of over 5 years (issue 2011) all articles which included the term *replicat** in the text were analyzed, taking into account the complete history of each publication.

The overall rate of replication found in the study was 0.13% (221/164,589 articles), eight times less than that estimated in Psychology by the same authors (Makel, Plucker & Hegarty, 2012). A positive evolution of this over time was also observed. Since 1990 the rate has been almost four

times higher than this. In other words, replication studies have increased from around 1 in every 2000 articles to almost 1 in every 500.

Regarding the analysis of results, the general rate of successful replications was around 67.4%. Although these results, which suggest that most studies were successfully replicated, differed from those recorded for Medicine or the Health Sciences (being significantly less promising), they appear to be similar to rates recorded in Psychology. Nonetheless, the authors call to mind the fact that almost half of the replications (48.2%) were carried out by the same teams that had published the original study. Moreover, the replication rate was higher when performed by the same team or when there was some degree of overlap in the authorship than when conducted by a totally independent team. It is important, therefore, to consider possible limitations associated with self-replications. It is uncertain whether the higher success rates arise from greater knowledge and experience or if the biases and questionable research practices recorded in the published literature of the Social Sciences are also replicated, a question that requires to further investigation.

In any case, it appears that replication is rare and should be encouraged. As the authors of this work explain, replication is not the panacea and will not resolve all the challenges and problems related to the rigor, reliability, precision and validity of educational research. However, to continue to ignore it, either implicitly or explicitly, suggests a deep lack of understanding about science and how to further its advancement.

These considerations have extended throughout the area of Psychology, where the crisis we describe here has been profound and has specifically developed in the past few years in the phenomena of replication and reproducibility. We briefly mention here the previously cited work of the *Open Science Collaboration,* published in *Science*. This article had, and still has, an enormous impact (1225 citations in Google Scholar by the 1st September 2017), and has promoted an interesting academic debate attracting both criticism and counter-criticism that has turned this study into an essential reference work in barely two years.

It is an ambitious, collaborative and large-scale project in which independent research teams from all over the world participate to obtain an initial estimate of the reproducibility of Psychology research. More specifically, replications were performed of 100 correlational and experimental studies published in 2008 in three high impact Psychology journals (*Psychological Science*, *Social Psychology*, *Experimental*

Psychology: learning, memory and cognition) using high power designs and original materials where available. A total of 5 indicators were used to evaluate the success of a replication: significance and p-value, effect sizes, subjective assessment of the replication teams and meta-analysis of the effect sizes. The results of these five indicators are complex and difficult to summarize in a few paragraphs as they also include the analysis of possible factors related to reproducibility and exploratory studies were carried out with the aim of inspiring future research. However, some of the data are highly illustrative. A total of 97% of the original studies reported significant results, while results were only significant in 37% of the replications. The mean effect size estimated in the replications was around 0.197, half of that reported in the original studies (0.403), corresponding to a substantial decrease. Finally, 39% of the effects were considered subjectively to be replications by the research teams.

Taken together, the authors have drawn clear conclusions from these and the remaining results. A large proportion of the replications show weak support for the original results in spite of using the same material as that provided by the authors, reviewing beforehand the methodological reliability and the high statistical power to detect the original effect sizes.

In any case, as the authors point out, this work leaves many questions unanswered and should be considered as a first step and not as the point of arrival. The original results offer tentative evidence, and the replications add additional confirmatory evidence. In some cases, the replications increase confidence in the original results. In other cases, however, they suggest that more research is required to establish the validity of the original results.

Open Science

Unanimous agreement about the need to increase transparency and openness in the processes used to carry out, disseminate and assess scientific research (as key factors to guarantee reproducibility) has led to the design and implementation in recent years of a range of initiatives, which together have given rise to what is known as *Open Science* (Morey et al., 2016; Nosek et al., 2015). An overall view and a good illustration of the kind of initiatives that are emerging are those promoted by the ***Center for Open Science*, a North American non-**

profit organization aimed specifically at developing free of charge and freely accessible technological tools for the promotion of open science and reproducibility (https://cos.io/**). Here, we describe three of its projects that are particularly representative of developments in this field.**

Guidelines for Transparency and Openness Promotion- TOP Guidelines

This proposal includes standards for eight modules that can be either partially or totally adopted and developed by scientific journals. Each module also includes three stringency levels for transparency. A flexible framework is defined that acknowledges that not all the standards are applicable to all journals or disciplines. The modules correspond to:

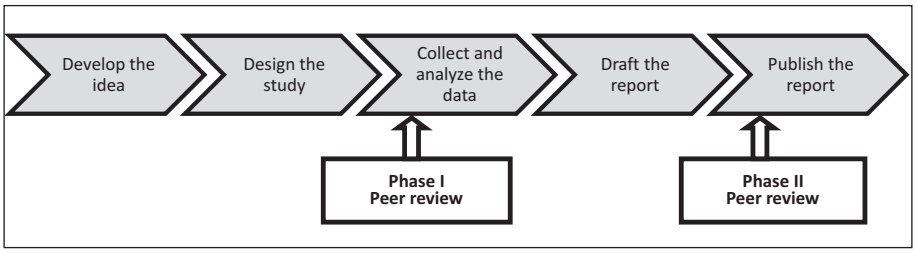
- Citation standards
- Data transparency
- Transparency of analytical methods (code)
- Transparency of research materials
- Transparency of analysis and design
- Pre-registration of studies
- Pre-registration of analytical plans
- Replication

An increasing number of journals are adopting this framework in a variety of disciplinary fields. For example, the journal *Nature* and those published by Springer and Biomed Central joined on 15th March 2017.

We regard the concept of preregistered articles and reports to be particularly interesting. This is a new publication format that emphasizes the importance of the research question and the quality of the methodology by peer reviews prior to data collection (Nosek & Lakens, 2014). This entails provisionally accepting high quality research protocols that will be published regardless of the results obtained if the authors follow the specified methodology (see graph II).

Received: 09-04-2017 Accepted: 23-02-2018

GRAPH II. Preregistration of articles and analysis plans



Source: Center for Open Science (https://cos.io/rr/)

The format is designed to encourage good practices and to do away with a series of questionable research practices, such as low statistical power, selective reporting or publication bias, while it enables flexibility to complete the report with unpredicted results, qualifying them as such. Emphasis is placed on the distinction between exploratory and confirmatory research to guarantee a correct assessment and evaluation of the results of the scientific research.

At the time of writing this article, more than one hundred journals use this format, either including it as a standard part of their acceptance policy for original manuscripts, or as part of special editions.

Open Science Framework - OSF

Open science framework is an open and free of charge *web* environment that enables one to connect with and support the work flow of the research over the entire duration of the project, which can be used to collaborate, document, store, share and record research projects, materials and data. Other applications and platforms exist, but this is a good example of the types of tools and innovative software developed to support new practices of open science. Some of the journals that recommend using this platform as a repository to share data and materials and to preregister articles are: *Biomed Central journals, Cognition, Nature, Psychological Science, Perspectives on Psychological Science, PloS, Science, Scientific Data*.

Open Science Badges

These are used to expressly identify and recognize articles that comply with certain standards of transparency and openness. They can be granted after a simple declaration made by the authors or as the result of a peer review process, depending on the editorial policy of each journal.

On writing this article around twenty journals had adopted these badges. For instance, requirements to be awarded an open data badge include:

- Providing a URL, DOI, or another permanent *path* for accessing the data in a public open-access repository.
- To provide together with the data sufficient information for an independent researcher to be able to reproduce the results reported in the article (metadata).
- To have an open type license that allows others to copy, distribute and use the data, while maintaining the owners' rights and credit where necessary (*Creative Commons* has recently defined several licenses of this type).

On the 1st August 2017 the American Psychological Association – APA announced an agreement with the *Center for Open Science* by which its journals offer open science badges and will use the Open Science Framework (OSF) as a data repository and to manage the preprint services (APA, 2017).

Conclusions: some challenges for the future of Spanish educational research

From the review conducted in the previous sections it is important to attempt to define some challenges that could be faced and must be overcome by educational research in our country. Here, we separate these into four different dimensions.

Received: 09-04-2017 Accepted: 23-02-2018

An analysis of current practices

When preparing this study we found there to be no strong tradition in reviewing and assessing educational research practices in Spain, unlike in other contexts, or in ours but in different areas. For example, in the area of Psychology studies can be found that analyze the use of statistical methods to identify good and bad practices and to formulate usage recommendations (in addition to the cited works of Caperos & Pardo, 2013 and Garcia, Campos & de la Fuente (2011); also see that of Izquierdo, Olea and Abad, 2014, among others). We can perhaps identify here a first question that must be addressed. We do not currently have sufficient information or evidence to enable us to diagnose, and where necessary to correct, incorrect practices. Therefore, a necessary and timely line of work would be to analyze the current statistical and methodological practices in Spanish educational research.

Training, informing and increasing awareness of researchers

Given the high prevalence with which questionable research practices have been identified, it seems reasonable to ask ourselves if, in fact, the researchers are always fully aware of engaging in them. Considering this possibility would perhaps help us to understand the situation better. Some of them may underestimate or be unaware of the complexity of some of the statistical techniques, for example, they may have difficulty understanding the impact of questionable research practices on type I error rates. We must focus on intentionality to distinguish malpractice (from a lack of information or insufficient training) from fraud. As explained in the literature (Sijtsma, 2016), there may be a need to change the initial and on-going training processes of researchers in statistics and methodology. Now is perhaps the moment to reflect upon initial postgraduate training and also upon the opportunity to arbitrate more permanent training approaches for active researchers.

Moreover, the difficult and slow adoption of innovations and new good practices by researchers is receiving an increasing amount of attention (Henson, Hull & Williams, 2010; Sharpe, 2013; Cohen, 2017) and we believe that this issue should also be addressed in our setting.

Editorial policies: the role of editors and reviewers

At the end of the day, researchers want to publish their work and should be willing to comply with the demands of scientific journals. Initially, therefore, the journals play an essential role in the way in which research is conducted and disseminated. It would also be reasonable to deduce that if the journals explicitly impose adequate statistical and methodological standards for the acceptance of original manuscripts, which are accepted and considered as a standard of reference by the reviewers, and do not penalize or invite the publication of replications and encourage open science policies then this would have a direct positive effect, as appears to be occurring in other settings (see Munafò et al., 2017). In practice, we consider this to be a fantastic challenge that has arisen in our area and setting. During this decade in our country we have presided over an extraordinary consolidation of a large number of quality scientific journals in Education, witnessing a great and often selfless effort by part of the research community, and especially of editors and reviewers (Ruiz-Corbella, Galán & Diestro, 2014). A large number of publications have entered international indices and we personally consider that just to maintain previous levels of achievement would already be a challenge in itself. Nonetheless, we believe that these lines of action point towards a necessary way forwards for the future, a journey already embarked upon by some prestigious journals of education (E.g. see Lopez, Valenzuela, Nussbaum & Tsai, 2015).

Academic and scientific policies: the role of agencies funding and assessing research production

Last, but undoubtedly not least, we believe it is essential to consider the key role of the agencies that finance and assess the scientific production of researchers, teams and institutions, because, ultimately, they are the ones that design the incentive and recognition systems that greatly modulate individual and group behavior. In the assessment of production if quantity is awarded above all else and the ecosystem in which researchers work is mainly dominated by the pressure to publish, then it will be difficult for some of the practices described here to improve. If openly publishing data is not actively encouraged in any way then it is unlikely to become

Received: 09-04-2017 Accepted: 23-02-2018

more widespread. Moreover, if participation in collaborative broader scope projects and research consortia are not encouraged this will hinder their introduction and consolidation in practice. We agree, therefore, with authors who have emphasized the importance of introducing changes at this institutional level (Smaldino & McElreath, 2016), which is where most of the challenges described throughout this article are concentrated.

One final reflection

We essentially aim to emphasize in this work that scientific progress is an accumulative process to reduce uncertainty (rather than to obtain the right answers), and that this process can only be successful if the science itself is based on a natural and systematic scepticism about its own findings. We consider that the analysis carried out clearly shows that there is room for improvement in scientific practices. However, we hope we have also shown that, faced with the very plausible hypothesis that the current environment of science and scientific culture can be negatively affecting the validity and reproducibility of its results, the scientific system is behaving as it should do: systematically and critically reviewing its own practices and designing new methods of self-correction. We consider that a perspective like the one proposed here could help us to reflect on our own practices and, ultimately, contribute to the advancement of educational research in our country.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. doi:10.3102/0013189X035006033.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC:

American Psychological Association.

American Psychological Association. (2017, August 1). *APA Journals Program collaborates with Center for Open Science to advance open science practices in psychological research*. Retrieved from: http://www.apa.org/news/press/releases/2017/08/open-science.aspx.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452-454. doi:10.1038/533452a

Balluerka, N., J. Gómez, et al. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 1(2): 55-70.

Caperos, J.M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414.doi: 10.7334/psicothema2012.207.

Cohen, B. H. (2017). Why the resistance to statistical innovations? A comment on Sharpe (2013). *Psychological Methods, 22*(1), 204-210. doi: 10.1037/met0000058.

Cumming, G. (2014).The new statistics: why and how. *Psychological Science, 25*,7–29.doi:10.1177/0956797613504966.

Fernández-Cano, A., & Fernández-Guerrero, I. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*. Madrid: La Muralla.

García, J. Campos, E., & De la Fuente, L. (2011). The use of the effect size in JCR Spanish journals of Psychology: from theory to fact .*The Spanish Journal of Psychology*, *14*(2), 1050-1055.

Goodman, S.N. (2016). Aligning statistical and scientific reasoning. Misunderstanding and misuse of statistical significance impede science. *Science, 352* (6290), 1180-1181. doi: 10.1126/science.aaf5406.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (2016). *What if there were no significance tests?* Classic Edition. New York: Routledge.

Henson, R.K., Hull, D.M., & Williams, C.S. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educational Researcher, 39*(3), 229-240. doi: 10.3102/0013189X10365102.

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLoS Biology*, *13*(10),1-7. doi: 10.1371/journal.pbio.1002264.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences:

detection, prevalence, and prevention. *Trends in cognitive sciences, 18*(5), 235-241.

Izquierdo, I., Olea, J., and Abad, F.J. (2014). Exploratory Factor Analysis in validation studies: uses and recommendations. *Psicothema, 26*(3), 395-400.

Ledgerwood, A. (2014). Introduction to the Special Section on Advancing Our Methods and Practices *Perspectives on Psychological Science,* 9(3) 275–277. doi: 10.1177/1745691614529448.

Lopez, X., Valenzuela, J., Nussbaum, M. & Tsai, C.(2015). Some recommendations for the reporting of quantitative studies (Editorial). *Computers & Education, 91*,106-110.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304 –316. doi: 10.3102/0013189X14545513.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542. doi:10.1177/1745691612460688.

McNutt, M. (2014). Journals unite for reproducibility (Editorial). *Science, 346*(6210), 679-679. doi:10.1038/515007a.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming (2014). *Psychological science*, *25*(6), 1289-1290.

Morey, R.D. et al. (2016). The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science, 3*, 150547. doi: http://dx.doi.org/10.1098/rsos.150547.

Mulder, J. & Wagenmakers, E.J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology, 72*, 1–5.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. doi:10.1038/s41562-016-0021.

Nature (2014). Journals unite for reproducibility (Editorial). *Nature, 515*(7525), 7.

Nosek, B. A., et al. (2015). Promoting an open research culture. *Science, 348*, 1422-1425. DOI: 10.1126/science.aab2374.

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

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

Peng, C-Y., Chen, L-T., Chiang, H-M., & Chiang, Y-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*, 157-209. doi: 10.1007/s10648-013-9218-2.

Peng, R.D. (2015). The reproducibility crisis in science. A statistical counterattack. *Significance*, 30-32

Perspectives on Psychological Science. (2012). Special section on replicability in psychological science: A crisis of confidence? Retrieved from: http://pps.sagepub.com/content/7/6.toc

Perspectives on Psychological Science. (2014). Special section on Advancing our methods and practices. Retrieved from: http://journals.sagepub.com/toc/ppsa/9/3.

Ruiz-Corbella, M., Galán, A. & Diestro, A. (2014). Las revistas científicas de Educación en España: evolución y perspectivas de futuro. *RELIEVE, 20* (2), art. M1. doi: 10.7203/relieve.20.2.436.

Savalei V., Dunn E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology, 6*, 245. doi: 10.3389/fpsyg.2015.00245.

Schweinsberg, M. et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55-67. doi: https://doi.org/10.1016/j.jesp.2015.10.001.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*(4), 572. doi: 10.1037/a0034177.

Sijtsma, K. (2016). Playing with data–Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*(1), *1-15*. doi:10.1007/s11336-015-9446-0.

Sijtsma, K., Veldkamp, C.L.S. & Wicherts, J.M. (2016). Improving the conduct and reporting of statistical analysis in Psychology. *Psychometrika, 81*(1), 33-38. doi:10.1007/s11336-015-9444-2.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis

allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi:10.1177/0956797611417632.

Smaldino, P.E., McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 160384. doi:http://dx.doi.org/10.1098/rsos.160384.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes (pp. 246-262). En Osborne, J.W. (2008)(Ed.). *Best Practice for Quantitative Methods*. London: Sage.

Trafimow, D. & Marks (2015). Editorial. *Basic and Applied Social Psychology, 37* (1), 1–2.

Waldman, I. D., & Lilienfeld, S. O. (2015). Thinking about data, research methods, and statistical analyses: Commentary on Sijtsma's (2014) "Playing with data". *Psychometrika*, *81*(1), 16-26. doi:10.1007/s11336-015-9447-z.

Warschauer, M., Duncan, G. J., & Eccles, J. S. (2015). Inaugural Editorial: what we mean by "open". *AERA Open,1* (1),1-2. doi: 10.1177/2332858415574841.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician, 70*(2), 129-133. doi: 10.1080/00031305.2016.1154108.

Wigboldus, D. H. J., & Dotch, R. (2015). Encourage playing with data and discourage questionable reporting practices. *Psychometrika, 81*(1), 27-32. doi:10.1007/s11336-015-9445-1.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594.

Información de contacto: Ángeles Blanco Blanco. Universidad Complutense de Madrid, Facultad de Educación, Departamento de Investigación y Psicología en Educación. Rector Royo Villanova s/n 28040 Madrid. E-mail: ablancob@ucm.es