REVISTA CUBANA DE CIENCIAS BIOLÓGICAS
http://www.rccb.uh.cu

ARTÍCULO DE REVISIÓN

Las crisis actuales de la Ciencia

Current crisis in Science

Dennis Denis

Departamento de Biología Animal y Humana, Facultad de Biología, Universidad de La Habana, Cuba.

Autor para correspondencia: *dda@fbio.uh.cu*

RESUMEN

Nunca antes en la historia de la Ciencia y la Tecnología se había dado un escenario tan complejo como el que existe en la actualidad. Los avances tecnológicos, la informatización, las presiones económicas, incertidumbres políticas y cambios sociales y ambientales a escalas globales, tienen efectos profundos en la Ciencia y han hecho aflorar un conjunto de problemas que han sido reconocidos con el nivel de crisis. Estos son la crisis de la competitividad y sus efectos, sobre todo los fraudes, conductas científicas inadecuadas y las llamadas Prácticas Cuestionables de Investigación; la crisis de la estadística, la crisis de la credibilidad y la crisis de la reproducibilidad y replicabilidad. Si bien se ha sospechado que estos problemas siempre habían existido, en los últimos años un amplio conjunto de publicaciones han demostrado que su prevalencia alcanza niveles extraordinarios y para combatirlos se están comenzando a implementar cambios fundamentales en la manera en que se hace, reporta y evalúa la propia Ciencia.

Palabras clave: competitividad, credibilidad, reproducibilidad, conductas científicas inadecuadas, ética

ABSTRACT

Never before, in the history of Science and Technology there had been such a complex landscape as the one that currently exist. Technological breakthrough, informatization, economic pressures, political uncertainties and social an environmental changes at global scales are having profound effects on Science and a series of negative issues had arisen and recognized as real crisis. Those are competitiveness crisis and its effects, mainly fraud, scientific misconduct, and the named Questionable Researches Practices; statistical crisis, credibility crisis and reproducibility and replicability crisis. Been suspected for a long time, these problems had emerged in last decades, and a wide group of papers have proven that its prevalence reaches extraordinary levels. To fight them, some fundamental changes are been implemented in the ways of making, reporting and assessing Science.

Keywords: *competitiveness, credibility, reproducibility, misconduct, ethics*

Recibido: 2020-04-20

Aceptado: 2020-05-28

INTRODUCCIÓN

Nunca antes en la historia de la Ciencia y la Tecnología se había dado un escenario tan complejo como el que existe en la actualidad. Los avances tecnológicos han desembocado en una generación de datos sin precedentes, entre presiones económicas, incertidumbres políticas, así como cambios sociales y ambientales a escalas globales. En este contexto, se están produciendo profundas transformaciones en la manera en que se desarrollan e intercambian los resultados científicos.

Las tendencias actuales en las ciencias de la vida, como en toda la Ciencia, están siendo moldeadas por dos factores fundamentales: el impacto de la informatización y los problemas éticos que han aflorado con la modernidad. Dentro de estos problemas, hay cuatro fundamentales que, por su magnitud e impacto, se han calificado al nivel de verdaderas crisis, por las revistas e instituciones científicas más importantes del mundo. En la lucha contra ellas, se están generando cambios fundamentales en la manera de actuar de los científicos y de comunicar la Ciencia, así como en su relación con la sociedad.

Estos cuatro problemas, aunque interrelacionados, se reflejan en diferentes aspectos centrales a la actividad científica y son: el problema de la competitividad entre investigadores y sus efectos, la pérdida de credibilidad científica, la crisis de la estadística y la crisis de reproducibilidad y replicabilidad. Muchos de estos, sino todos, son problemas que siempre han existido. En 1830, Charles Babbage en una publicación titulada "*Reflections on the Decline of Science in England*" se quejó de la generalización de prácticas nocivas a la investigación que llamó: "*hoaxing, forging, trimming and cooking*" ("engaños, falsificaciones, adornos y cocinados") (citado por Broad y Wade, 1982). Pero en la actualidad, el ritmo acelerado de la Ciencia y el advenimiento de grandes avances relacionados con informatización: internet, metabuscadores y programas capaces de revisar y comparar gigantescos volúmenes de publicaciones en poco tiempo, e identificar casos de fraudes, plagios, repeticiones y otros aspectos negativos, han hecho muy evidente la extensión de estos problemas.

En esta revisión se desea traer a la discusión estos temas relacionados con aspectos éticos, metodológicos y de transparencia científica, al contexto de la comunidad científica cubana relacionada con las ciencias de la vida. Aunque sea incómodo, aceptar públicamente y hablar sobre los problemas que existen permite reflexionar sobre cuáles son las prácticas que se deben potenciar y cuales evitar, tanto desde el punto de vista de la educación superior, como desde el seno de las revistas académicas y direcciones institucionales relacionadas con la política científica.

LA CRISIS DE LA COMPETITIVIDAD Y SUS EFECTOS DERIVADOS

El primer problema de la Ciencia, el de la competitividad, ha emergido como producto de la cultura "publica o perece", resultante del empleo generalizado de criterios inapropiados para evaluar la productividad y la calidad de los investigadores. Los investigadores son evaluados rutinariamente por los empleadores, tribunales o comités científicos diversos para decidir las categorizaciones – con los aumentos salariales correspondientes, las titulaciones o el simple desempeño laboral durante el año, que determina a veces la recepción de algún premio o estímulos monetarios. El enfoque casi absolutista de medir el mérito científico por el número de publicaciones y el factor de impacto de las revistas donde son ubicadas ha generado un escenario donde hay que publicar en estas para sobrevivir.

Esto se ha empeorado con la aparición de revistas oportunistas con bajos umbrales de calidad, que se han llamado "factorías de artículos" o "revistas depredadoras", que compiten entre sí por publicar la mayor cantidad de trabajos. Estas emplean múltiples estrategias de captura de autores, por medio de correos *spam*, bajando las tasas de pago por artículos, acelerando los tiempos de publicación (Kearney, 2015; Eriksson y Helgesson, 2016).

Como consecuencia de la cultura de *publish or perish* han derivado otros problemas como la degradación del concepto de autor para promover y justificar la multiautoría (Hernández-Chavarría, 2007; Denis, 2017) o un brote casi epidémico de conductas científicas inadecuadas (Kaiser, 1995; Abbott *et al*., 1999). Estas son mucho más generalizadas de lo que generalmente se está dispuesto a reconocer dentro de la comunidad científica, y de lo cual muchas veces se evita hablar (Arst, 2000), a pesar de que afecta fuertemente campos tan sensibles como el de la medicina y la salud humana (Lock *et al.,* 2001).

Los peores casos, relacionados con fraudes, son más frecuentes de lo que la mayoría de los científicos desean creer (Hartemink, 2000) al confiarse inocentemente en una honestidad utópica. Los miles de casos destapados de fraudes científicos han creado un ambiente de recelo y escepticismo hacia la Ciencia, no solo desde la opinión pública mundial sino entre los propios investigadores.

Se han divulgado casos tan vergonzosos como el artículo de Steinschneider (1972) sobre el vínculo de la apnea con el síndrome de muerte súbita infantil. Este caso derivó en años de investigaciones y millones invertidos en estudios biomédicos, que también terminaron de muerte súbita con el descubrimiento de que la publicación había sido resultado del asesinato confeso de los cinco niños por la madre, Waneta Hoyt (Crimson, 1994).

También destaca el escandaloso caso Hwang, del científico coreano que falsificó los resultados de investigaciones en líneas celulares humanas publicados en revistas de punta (Fig. 1).

Todos estos eventos han hecho tambalearse la imagen de credibilidad científica (Delgado-López-Cózar *et al*., 2007). Y no son solo artículos aislados, existen muchos casos donde se ha demostrado la publicación de decenas o incluso centenares de artículos fraudulentos por un mismo investigador (Schulz y Katime, 2003).

Un escándalo relativamente reciente se dio a la luz en noviembre de 2011, cuando Diederik Stapel, psicólogo de la Universidad de Tilburg en Holanda, fue investigado y confesó haber falsificado los datos de 30 publicaciones de alto impacto en su campo (Tilburg University, 2011). Posteriormente, en un libro autobiográfico titulado "*Faking Science: A True Story of Academic Fraud*" este autor explicó las causas que lo llevaron a estos hechos (Stapel, 2012).

Pero no se desea profundizar en el aspecto del fraude "crudo", que ya está recibiendo atención fuerte en las políticas científicas internacionales. Por ejemplo, la revista *Nature* ha dedicado varios volúmenes a su discusión.

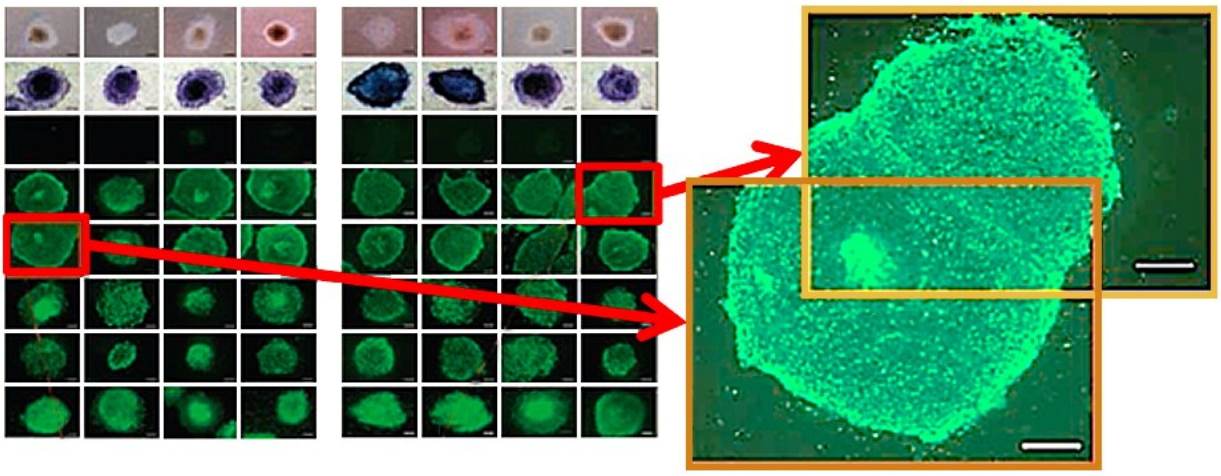

Figura 1: Evidencia de manipulación de las fotografías en los materiales suplementarios del artículo de Hwang *et al.* (2005), publicado en *Science* y retractado luego de la demostración de la fabricación de los datos (fuente K. Buckheit/ *Science;* https://science.sciencemag.org/content/311/5757/news-summaries). La imagen muestra como fotografías reportadas como líneas celulares diferentes son solo diferentes secciones de una misma foto.

Figure 1: *Evidence of manipulation in pictures presented as supplementary materials of the paper of Hwang et al. (2005), published in Science and retracted after probed data manufacturing (source K. Buckheit/Science; https:// science.sciencemag.org/content/311/5757/news-summaries). The plate shows pictures reported as different cell lines actually are just sections of the same photograph.*

Es necesario hablar de otros problemas más "benignos" que se llamado en la literatura "Prácticas Cuestionables de Investigación" (QRP del inglés *Questionable Research Practices*). El término QRP, popularizado por John *et al*. (2012), agrupa una gran diversidad de procederes (Tabla 1) y que, como epidemias emergentes de enfermedades históricas conocidas, se han estado extendiendo en el seno de la actividad científica moderna (Fiedler y Schwarz, 2016). Entre estas, el *HARKing*, las bajas potencias, el hackeo de la p y los sesgos de publicación han sido llamados los "*cuatro jinetes del apocalipsis científico"* (Fig. 2).

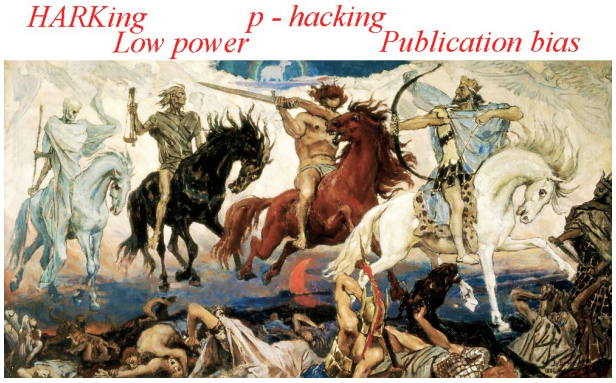

Figura 2. Representación popularizada en varios eventos científicos de cuatro de las principales Prácticas Cuestionables de Investigación. Sobre el cuadro Apocalipsis, 1887 del artista ruso Viktor Vasnetsov.

Figure 2. *Popularized representation of four main Questionable Research Practices in several scientific meetings. Using Apocalypses, 1887 painted by the Russian artist Viktor Vasnetsov.*

Encuestas sobre la extensión de estas prácticas han dado resultados alarmantes dentro de la comunidad de psicólogos (John *et al.,* 2012; LeBel *et al.,* 2013; Agnoli *et al.,* 2017), donde incluso han sido defendidas como válidas por un elevado porcentaje de investigadores encuestados. Pero no es único de este campo: estudios en el campo de la ecología y evolución encontraron, posteriormente, prevalencias comparables (Fraser *et al.,* 2018).

Masterson (2018) informó que encuestas anónimas detectaron que, como mínimo, la mitad de los biólogos evolucionistas y ecólogos manipulaban rutinariamente los resultados para embellecerlos y hacerlos más publicables, surgiendo lo que O´Boyle *et al.* (2017) llamaron "efecto crisálida". Fraser *et al.* (2018) discutieron la amplia gama de prácticas cuestionables dentro de la investigación y demostraron - a través de 807 encuestas anónimas - que son actividades muy comunes dentro de estudios biológicos. Encontraron, por ejemplo, que el 64% de los investigadores reportaban haber dejado de publicar resultados por no haber encontrado significación estadística (*cherry picking*). El 51% confesó haber publicado resultados atractivos e inesperados como si hubieran sido hipótesis concebidas antes del estudio y puestas a prueba (*HARKing*) y el 42% reconoció haber recopilado más datos luego de haber analizado una muestra y darse cuenta de que había significación estadística (una forma de hackear la p).

LA CRISIS DE LA ESTADÍSTICA Y SUS MANIFESTACIONES

Varios casos entre estas prácticas cuestionables se relacionan con la segunda de las crisis actuales reconocidas en la Ciencia: la que tiene que ver con la estadística aplicada en los trabajos de investigación. Esta fue traída a la atención pública por el artículo de revisión *The Statistical crisis in science* publicado por Gelman y Loken (2014) en la revista *American Scientist*. Esta crisis no se relaciona, únicamente, con una aplicación errónea de métodos estadísticos, o a las críticas repetitivas del ignorar los problemas o sesgos de la estadística frecuentista clásica en detrimento del empleo de métodos alternativos más actuales. Además, incluye otras formas de actuación negativas desde el punto de vista metodológico.

La más conocida, ya mencionada entre las QRP, es el "hackeo de la p" que incluye varios aspectos que influyen en los resultados como la decisión de eliminación de valores atípicos (*outliers*) basándose en el impacto que tiene sobre las pruebas estadísticas o la toma de decisiones de muestreo guiándose por la significación lograda en una prueba (aumentar las muestras donde las pruebas iniciales dan resultados significativos o detener el muestreo cuando ya se logra la significación). También incluye el desarrollo de modelos donde la exclusión o inclusión de co-variables se guía por su efecto sobre el resultado principal (es decir, para "mejorarlo"). Simmons *et al.* (2011) demostraron, por medio de datos experimentales simulados, cómo esta práctica puede inflar las tasas de error por falsos positivos en la literatura. Un caso más burdo es el "redondeo de la p" (o sea, reportar una significación de p= 0,052 como p<0,05).

Esta apreciación distorsionada de la significación de un resultado hacia el limitado significado estadístico (p<0,05) conlleva al sesgo de publicación (otra de las QRP).

Tabla 1. Definición de las principales Prácticas Cuestionables de Investigación (QRP: *Questionable Research Practices*) y otros procedimientos no éticos relacionados con la publicación de resultados científicos que han emergido en la Ciencia actual.

Table 1*. Definition of the main forms of Questionable Research Practices (QRP) and other non-ethical procedures related to scientific result publishing that emerge in current science.*

Nombre *(nombre en inglés)*	**Definición**	**Fuentes**
Hipótesis *a posteriori* (HARKing: acrónimo de *Hypothesizing After Results are Known*)	Incluye la presentación de hipótesis luego de haber tenido los resultados o la presentación de resultados inesperados como si hubiesen sido predichos o esperados. Bajo este efecto, estudios que realmente son exploratorios (inductivos) se muestran como su fueran confirmatorios (hipotético –deductivos).	Kerr (1998)
Hackeo de la p (*p- hacking*)	Conjunto de prácticas de manipulación de datos que aumentan la probabilidad de obtener resultados estadísticamente significativos. Incluyen la eliminación sesgada de valores atípicos (*outliers*), la selección sesgada de datos, el aumento de las muestras luego de haber hecho las pruebas estadísticas.	Head *et al.* (2015)
Bajas potencias (*Low power*)	El empleo de muestras muy pequeñas, limitadas o con variabilidades muy grandes que conducen al fallo en la obtención de significación estadística pero que no prueban la inexistencia del efecto, aunque son interpretados así. Se expresa también por la sobre-confianza en estudios de baja potencia: sacar conclusiones sobreestimadas de muestras subestimadas.	Halsey *et al.* (2015) Button, *et al.* (2013)
Saturación estadística (*Significance excess*)	Uso exacerbado de pruebas estadísticas innecesariamente, para dar la imagen de rigor o profundidad, pero que conduce a la inflación de las tasas de error por experimento o investigación (inflación del alfa). Se relaciona con la tortura de los datos.	Parker y Nakagawa (2014)
Tortura de los datos (*data dredging o torturing*)	El empleo de los mismos conjuntos de datos para múltiples pruebas estadísticas, con el afán de detectar efectos que son poco evidentes. Se relaciona con lo que algunos autores llaman *data snooping*, *data fishing* o *data butchery*.	Davey y Ebrahim (2002)
Pseudorréplica (*pseudorreplication*)	Defecto del diseño de toma de datos que al ser analizados con pruebas de significación clásicas –que se basan en las varianzas- resultan en tasas infladas de falsos positivos, al considerarse independientes datos que no lo son (tomados sobre las mismas unidades o con distintas formas de autocorrelación –espacial o temporal). Se puede reconocer por una incorrecta identificación del tamaño de muestra o por confundirse el número de muestras con su tamaño.	Hulbert (1984)
Recolecta de cerezas (*Cherry picking)*	Se refiere a, dentro de un análisis de datos, la selección solo de aquellos resultados que han resultado significativos o estadísticamente "atractivos" para el reporte final. Es la actividad basal que conduce al sesgo de publicación.	Murphy y Aguinis (2019)
Sesgo de publicación (*publication bias*)	La publicación preferencial de estudios que obtienen resultados estadísticamente significativos, como si la no existencia de diferencias no fuera un resultado igual de válido. Se demostró su extensión desde el desarrollo de los estudios meta-analíticos. Es el resultado extremo o final del *cherry picking* en los análisis.	Thornton y Lee (2015)
Miopía de hipótesis (*hypothesis myopia*)	Fijación en una hipótesis de interés personal y enfocarse en buscar elementos de prueba e interpretar los resultados a su favor, sin buscar evidencias en contra u obviando que los mismos datos pueden apoyar otras hipótesis alternativas.	Nuzzo (2015)
Multiautoria injustificada	La inclusión en la lista de autores de personas que no han hecho lo suficiente para calificar como tales, para aumentar las tasas personales de publicación, lo que conlleva a que el número promedio de autores por artículo se haya incrementado en las últimas décadas.	Weltzin *et al*. (2006) Nayernouri (2009)

Aunque racionalmente, en el discurso oral, los científicos aceptamos la igual validez de cualquier resultado, tanto positivo como negativo, a la hora de publicar no somos consistentes y solo se tienden a publicar los positivos. En 1969, el estadístico T.D. Sterling encontró que 97% de las publicaciones en cuatro importantes revistas de Psicología solo reportaban resultados estadísticamente significativos. El estudio se repitió en 1995 y la proporción se mantenía (Sterling, 1969; Sterling *et al.,* 1995). Fanelli (2012) encontró que la proporción de publicación de resultados positivos aumentó en más de 22 % entre 1990 y 2007. Algunos autores identifican este problema con el término *File drawer*, en referencia al engavetado de resultados e investigaciones "no significativas" (Csada *et al.,* 1996). Esta es una evidencia de la extensión de la cuestionable práctica de "recolecta de cerezas" y que se hace muy marcada cuando se comparan los resultados de las tesis con los de las publicaciones que se derivan de ellas, en las cuales se duplica la proporción relativa de resultados significativos sobre los no significativos (O'Boyle *et al.,* 2017).

Las justificaciones dadas para el empleo de estas prácticas son las presiones para publicar, adaptándose al sesgo conocido de los editores y árbitros, ya que los trabajos que obtienen resultados significativos son "mejor vistos" y se publican más. También se llegan a aducir buenas intenciones, como el deseo de presentar una narrativa ordenada y coherente dentro de los artículos (Fraser *et al.,* 2018) evitando inseguridades o contradicciones internas. O sea, la presión por publicar puede considerarse como una fuerza selectiva importante que promueve el empleo de métodos que inflan la frecuencia de falsos positivos en los resultados de investigación. Smaldino y McElreath (2016) demostraron que estas prácticas también contribuyen a acelerar la diseminación de métodos pobres o inadecuados dentro de una comunidad de investigadores.

Finalmente, al debate de los problemas estadísticos en la Ciencia actual hay que traer un último elemento. El cálculo o interpretación de la p ha recibido denodadas críticas (Wasserstein y Lazar, 2016; Wasserstein *et al.,* 2019). Sin embargo, debe reconocerse que este es sólo el último paso del análisis estadístico y al final se reduce a una simple regla de decisión que ciertamente es muy fácil de atacar (Gelman y Stern, 2006).

Pero en la práctica, hay muchas decisiones previas que pueden tener más impacto aún en los resultados y que pocos se cuestionan: desde el diseño experimental, la manera en que se toman los datos, el cómo se controlan los factores confundidos, como se organizan y limpian las matrices de datos, cuál análisis exploratorio reciben, cómo se resumen o que tipos de modelos estadísticos se aplican o asumen. Los niveles de significación se alteran en dependencia del modo en que los datos fueron filtrados, resumidos o modelados (Simmons *et al.,* 2011). La p es solo la punta del iceberg, pero como plantearon Leek y Peng (2015): hay que cuidar que el resto no termine de hundir la Ciencia.

En un artículo en ScienceNews, Siegfried (2014) dijo que las pruebas estadísticas para comprobar hipótesis tenían más fallos que las políticas de privacidad de Facebook. A esto respondió Leek (2014) que el problema no era que las personas usaran mal los valores de p, sino que la vasta mayoría de los análisis de datos son ejecutados por personas que no están entrenadas de forma adecuada. Kaelin (2017) se lamentaba de que la investigación científica había cambiado, de buscar la comprobación de conclusiones estrechas por múltiples vías, a hacer aseveraciones muy amplias y generales basadas en evidencias limitadas - como las aportadas por las pruebas de significación estadística: los artículos se parecen cada vez más a mansiones de paja que a pequeñas casas de ladrillos. Amrhein *et al*. (2019) presentaron en *Nature* una solicitud avalada por más de 800 signatarios para pedir a todas las revistas eliminar el concepto de significación estadística, pero siendo objetivos, la comunidad científica muy probablemente continuará ignorando las razones y la crisis estadística continuará.

LA CRISIS DE LA PÉRDIDA DE CREDIBILIDAD

Si sumamos el aumento de la competencia por publicar, la disminución de los umbrales de calidad de muchas revistas y la extensión demostrada de los fraudes y prácticas cuestionables, se llega a la tercera de las crisis globales que enfrenta la Ciencia: la pérdida de credibilidad. ¿Cómo es posible que exista un movimiento que no es pequeño, incluso a nivel de gobiernos, que niegan la existencia y efectos del Cambio Climático? ¿Cómo es posible que ante cada nueva emergencia de una nueva enfermedad, la primera sospecha recae sobre científicos que pudieron haberla creado en los laboratorios?

Es fácil culpar a los intereses económicos y políticos, pero pocos investigadores se atreven a considerar siquiera la propia responsabilidad de la comunidad científica en la falta de confiabilidad actual en sus opiniones o resultados. Fiske y Dupree (2014) publicaron un análisis de encuestas públicas que demostraban que los científicos ya eran vistos por la opinión pública como personas competentes pero poco confiables. Barr (2014) en la sección de Ciencia del sitio KQED (https://www.kqed.org/science) analizaba las causas de esta opinión generalizada y divulgada inconscientemente por los medios (Fig. 3) y sus consecuencias.

Figura 3. Imagen de los legos (LEGO®) de científicos locos, que son comercializados para niños, que emplea Barr (2014) como encabezado de su publicación, ya que contribuyen a la imagen negativa que los medios dan a los científicos fomentando la desconfianza hacia ellos. En el pie de la figura hace el comentario: "*No wonder the public doesn't trust scientists. Who'd trust these guys*?" (Imagen de flickr.com)

Figure 3. *Image of mad scientist Legos (LEGO®), sold to children, used by Barr (2014) as banner in his publishing, as example of things that contribute to the negative image of scientists that generate untruthfulness to them. In the foot he comments: "No wonder the public doesn't trust scientists. Who'd trust these guys?" (Image from flickr.com)*

Wooster (1998) hizo un llamado que, a pesar de aparecer en la revista *Science*, fue ignorado: el peligro asociado a que los investigadores que producen los resultados científicos sean los mismos que los divulguen popularmente en los medios u otros contextos de decisión. Mucho se ha hablado sobre la responsabilidad que tienen los científicos en la divulgación pública o popular de sus resultados, sobre todo bajo la asunción de que son quienes mejor conocen el tema y los que tiene la mayor credibilidad pública (más que los periodistas o presentadores televisivos).

Sin embargo, hay una diferencia fundamental entre publicar en Ciencia, que supone incluir los hallazgos derivados de estudios planificados, sistemáticos y objetivos, y divulgar "opiniones" sobre la relevancia o importancia de un resultado, lo cual involucra puntos de vista personales y subjetivos. Mills (2000) retomó el tema, en asuntos de políticas ambientales, y aclaró que la tarea del científico es informar con objetividad a los tomadores de decisiones, pero no debería ni tomar él las decisiones ni involucrarse personal o emocionalmente en ellas, ya que eso implica un evidente conflicto de intereses. Wooster (1998) mencionó que estas situaciones lo ponían nervioso, por la dificultad en diferenciar lo que es reportar de forma objetiva y lo que es predicar.

Si un científico ecologista hace un planteamiento público apocalíptico o exagerado con el objetivo de promover concientización y movilizar la opinión pública hacia un sentido, está haciendo un intercambio: está arriesgando o sacrificando su credibilidad por su apasionamiento. El fin no justifica los medios y no hay forma de saber las repercusiones que traerá esta pérdida generalizada de confianza en la palabra de los científicos. La credibilidad es uno de los principales pilares de la Ciencia que debería ser defendido a toda costa. Pero como resultado de los problemas que están emergiendo en la actualidad este pilar se está erosionando.

Los problemas asociados a los fraudes, las conductas científicas inadecuadas que están aflorando producto de la competitividad y los problemas estadísticos y metodológicos han llevado a que Ioannidis (2005) en *PLOS medicine* hiciera la grave y contundente afirmación de que la mayoría de los resultados de investigación publicados son falsos. El Editor en jefe de la revista *The Lancet* también llegó a decir públicamente: "*much of the scientific literature, perhaps half, may simply be untrue*" (Horton, 2015).

Y la pérdida de credibilidad no es solo desde fuera de la Ciencia sino, incluso dentro de la propia comunidad de científicos. Es reconocido que aunque se promueve la objetividad, es inevitable que exista un alto grado de subjetividad en el reporte de los resultados de las investigaciones. Lo que un investigador percibe como importante, los trabajos previos que selecciona para el encuadre de sus resultados, la explicaciones que ofrecen a patrones observados y el juicio sobre el valor potencial de sus hallazgos son aspectos netamente subjetivos. Pero a diferencia del público no

especializado, los lectores de los artículos son otros científicos, entrenados para reconocer y evaluar estas opiniones (una de las razones fundamentales por las que se separaron los acápites de Resultados y de Discusión) y escépticos por naturaleza.

Por ello, la pérdida de credibilidad dentro de los científicos no deriva de esta subjetividad sino de que existe la percepción de que la integridad científica es más débil de lo deseable. La mayoría de los científicos consideran que las prácticas cuestionables de investigación son usadas con alta frecuencia por otros colegas, tanto de sus propias organizaciones como de otras, independientemente de la edad o la experiencia de los investigadores (Fraser *et al.,* 2018). En algunos casos, la prevalencia sospechada que se ha expresado en las encuestas es mucho mayor que la autoreconocida, lo que sugiere que se realizan estas prácticas a pesar reconocerlas como socialmente poco aceptables. Incluso un 44,6% de los encuestados llegó a reconocer que tenían dudas sobre su propio uso o no de algunas de estas prácticas (Fraser *et al.,* 2018).

LA CRISIS DE LA REPLICABILIDAD Y REPRODUCIBILIDAD DE LAS INVESTIGACIONES CIENTÍFICAS

A todo lo anterior ha venido a sumarse un escándalo más reciente, que generó la cuarta de las crisis: el problema de las bajas replicabilidades y reproducibilidades que presentan los resultados científicos. Esta comenzó cuando en el año 2015, en la revista *Science*, se llamó la atención de que solo el 36% de los trabajos en el campo de la Psicología, al ser replicados lograban obtener los mismos resultados.

Shanks *et al.* (2015) demostraron como la distribución de los tamaños de efectos en los estudios publicados tendía a ser significativamente superiores a los que se obtenían cuando los trabajos eran replicados. Anteriormente, Forstmeier y Schielzeth (2011) y Button *et al.* (2013) habían mencionado el fenómeno *Proteus*, que se refiere a la manifestación de la "maldición del ganador" (*winner curse*). Este fenómeno aparece cuando en series de estudios que se desarrollan sobre un proceso o fenómeno, generalmente se demuestra que el primero es el más sesgado hacia un resultado extremo mientras que las réplicas posteriores tienden a encontrar tamaños de efectos menores o incluso contrarios. La crisis de replicabilidad no se restringe al campo de la Psicología sino que poco tiempo después se demostró en las ciencias biomédicas y ciencias biológicas en general (Schnitzer y Carson 2016; Schloss, 2018).

En todo este discurso se manejan tres términos que se confunden fácilmente, pero cuyo significado preciso debe tenerse claro (Kenett y Shmueli, 2015; Barba 2018). No es lo mismo que una investigación sea repetible, a que sea replicable o reproducible (Tabla 2).

Un experimento científico es repetible si los propios autores pueden hacerlo varias veces para confirmar los resultados y excluir la posibilidad de algún error de procedimiento o la intervención desafortunada del azar, o simplemente para aumentar el tamaño de muestra. Sin embargo, un estudio es replicable si otros autores pueden volver a hacerlo a partir de la información proporcionada en su publicación, pero en un nuevo contexto, con una nueva muestra e idealmente deberían llegar a las mismas conclusiones.

Tabla 2. Comparación entre los conceptos básicos manejados en el análisis de la crisis de replicabilidad y reproducibilidad de la ciencia en la actualidad.

Table 2*. Comparison among basic concepts frequently confounded used in analysis of replicability and reproducibility crisis in modern science.*

Concepto ***(Término en inglés)***	**Significado**	**Propósito**
Repetible *(Repeatibility - repeat)*	Si los autores pueden volver a realizar un estudio en exactamente las mismas condiciones en que se hizo el original, para verificar que se llega al mismo resultado.	Evaluar robustez
Replicable *(Replicability - replicate)*	Si otros autores pueden hacer un estudio similar a partir de la información proporcionada en la publicación de referencia, pero en un nuevo contexto y con una nueva muestra, con lo cual siempre aparecen ligeras variaciones.	Comparar con el original (evaluar credibilidad)
Reproducible *(Reproducibility - reproduce)*	Recrear los resultados y conclusiones por otras personas a partir de los datos primarios y los protocolos de análisis desarrollados y compartidos por los propios autores.	Verificar la corrección del análisis (evaluar confiabilidad)

La replicabilidad de los estudios es una piedra angular de la Ciencia. Ante la imposibilidad de saber con exactitud si un resultado es o no verdadero, el científico se apoya para dar solidez a sus conclusiones en el hecho de que sean replicables (filosofía inductiva). Por ello, el acápite de Materiales y métodos, donde se detalla el protocolo de obtención y análisis de los datos, tiene tanta importancia dentro de una publicación. Pero está claro que el tener que rehacer cada estudio para evaluar si es correcto o no es impracticable y poco eficiente en término de tiempo y recursos, así que históricamente los científicos hemos dado votos de confianza cuando leemos los resultados de otros investigadores. Los estudios meta-analíticos publicados en las últimas décadas (Fidler *et al.,* 2017) han generado muchas dudas sobre si esta decisión de confiar es o no acertada en la actualidad.

En el año 2016, la revista *Nature* condujo una encuesta enfocada a conocer el estado de opinión sobre este tema. Una muestra de 1576 investigadores de todo el mundo y todas las ramas de la Ciencia respondieron: 52% confirmó la existencia de esta crisis de replicabilidad, 38% la aceptó aunque "ligeramente" y solo un 3% respondió que no existía tal crisis (Baker, 2016). Más de la mitad de los encuestados reconoció que habían fallado en intentos de replicar investigaciones previas, 21% investigaciones propias y 31% estudios de otros… y el resto, no lo habían intentado nunca. Específicamente dentro de la Biología, 60% de los investigadores habían fallado en replicar estudios propios y más del 75% habían fallado en replicar experimentos de otros investigadores. En 2013, un proyecto norteamericano de 1,6 millones de dólares para evaluar la replicabilidad de investigaciones sobre el cáncer se propuso repetir 50 experimentos de publicaciones claves sobre esta enfermedad. Tuvieron que detenerse en 18, por no poder determinar exactamente cómo repetir los restantes estudios (Teytelman, 2018).

Esta crisis ha salido a los medios de divulgación masiva. Andrew Gelman publicó el 19 de noviembre de 2018 en *The New York times* un ensayo que se titulaba: "El experimento es fascinante. Pero nadie puede repetirlo". El haberse demostrado que los estudios cuando se replican o reproducen dan resultados diferentes a los publicados, en una proporción tan alta, contribuye fuertemente a crear dudas sobre la calidad de la Ciencia. Sobre todo cuando se ubica en el contexto de las críticas al uso de la estadística y a las prácticas inadecuadas, lo que hace que disminuya aún más la credibilidad.

Es cierto que la falta de replicabilidad de los estudios no es generalmente causada por conductas científicas inadecuadas conscientes sino que responde muchas veces a muchos otros aspectos más benignos (o menos malignos) - como por ejemplo hacer más énfasis, inconscientemente, en resultados llamativos que en otros detalles técnicos en los materiales y métodos. En este vital acápite de las publicaciones con frecuencia se ignoran aspectos básicos de los diseños como la forma de aleatorización empleada, los tamaños de muestra y su forma de cálculo, los tamaños de efecto, la presencia o no y el manejo de los valores atípicos, entre otros factores que imposibilitan una replicación precisa.

Se ha demostrado que, a veces, aspectos aparentemente muy poco importantes, como incluso la forma de mantener en cautiverio los animales de laboratorio, pueden producir variaciones importantes en los resultados que limitan su replicabilidad (Reardon 2016). Por otra parte, hay autores que conservan la vaguedad en los métodos empleados al publicar sus resultados para mantener una supremacía o ventaja competitiva sobre sus pares y no solo para ocultar deficiencias del trabajo o para aumentar su "publicabilidad" (Veld y Titus 2016), lo cual es éticamente muy cuestionable.

Para diferenciar los casos donde la ausencia de replicación está dada por la falta de información clave, de aquellos casos que se pueden replicar incluso con el protocolo detallado, Stark (2018) propuso el neologismo "pre-producibilidad". Un estudio es pre-producible si contiene toda la información detallada que permite su replicación.

Para poder replicar exactamente un estudio, generalmente se necesita mucha más información de la que aparece en la sección de materiales y métodos de un artículo: datos sobre los materiales usados (incluyendo los animales de laboratorio y su cuidado), instrumentos, procedimientos; diseño experimental, datos crudos salidos de los instrumentos; algoritmos usados para procesarlos; herramientas computacionales, incluyendo todos sus ajustes de parámetros o decisiones *ad hoc* tomadas durante el proceso; códigos, datos procesados; análisis intentados y desechados… De lo contrario, es muy probable que se falle en los estudios de validaciones o replicaciones por aspectos técnicos. En este sentido se están desarrollando alternativas como la inclusión de materiales suplementarios o la publicación de protocolos en sitios web especializados para ellos, y referirlos en las publicaciones.

La Ciencia avanza por medio de la corroboración de los resultados de un investigador por otros investigadores, y la obtención de nuevos aportes a partir de este punto. Esto asegura la consistencia general del cuerpo teórico de cualquier rama científica y su continuidad. La importancia de la repetición de los estudios fue enfatizada por Fisher (1935) en su clásico *The Design of Experiments*: "la confianza que se pone en un resultado no depende solo de la magnitud de los valores medios encontrados sino, por igual, en el grado de acuerdo entre experimentos paralelos". O sea, repetir estudios previos tiene un gran valor para la Ciencia pero lamentablemente estos son percibidos como poco importantes, y mucho menos si lo que se demuestra es que las diferencias entre resultados no son estadísticamente significativas – lo cual corroboraría la corrección del estudio previo. Esto nos regresa al problema de la crisis en la Estadística.

Por otra parte, hay que reconocer que hay estudios que por su propia naturaleza son sencillos de replicar o repetir (como los matemáticos, las modelaciones, trabajos metodológicos o trabajos experimentales simples) pero otros, en los que hay factores ajenos al control del investigador, pueden ser casi imposibles (Ives, 2018). Por ejemplo, un estudio con diseño BACI (*Before –After – Control – Impact*) del efecto del paso de un huracán por un ecosistema no puede ser evaluado por su reproducibilidad y ello no indica que sea incorrecto. Estos estudios se benefician más de un enfoque de triangulación que de replicación (Munafò y Smith, 2018). Sin embargo, la triangulación está fuertemente ignorada en la metodología de la ciencias biológicas.

En estos tipos de artículos de difícil replicación un estándar mínimo de aseguramiento de la calidad o veracidad de las conclusiones - entre la replicación total y no hacer nada - sería el poder reproducirlo (del inglés *reproduce*). Recuérdese que reproducir es recrear los resultados y conclusiones a partir de los datos primarios y de repetir el protocolo de análisis desarrollado, si estos son compartidos por los propios autores (Peng, 2009).

Sin embargo, la falta de reproducibilidad es el segundo de los componentes más preocupantes de esta crisis: la inmensa mayoría de los artículos científicos no son reproducibles (Rodríguez-Sánchez *et al.,* 2016). Esto deriva no solo de que los autores no dan acceso a los datos, o de que la escueta descripción verbal que aparece en la sección de métodos la mayoría de las veces es insuficiente para conocer todos los detalles del análisis (Ince *et al.,* 2012) o de los elementos de contexto de las investigaciones que no se pueden lograr de nuevo – sobre todo en el caso de los estudios de campo. El peor agravante de la crisis es que porcentajes inesperadamente altos de investigaciones incluso con los propios datos de los autores al ser reproducidos llegan a resultados diferentes. Por ejemplo, en genética poblacional, Gilbert *et al*. (2012) encontraron que el 30% de los artículos que utilizaban el paquete de r *structure*, cuando eran repetidos daban resultados diferentes a los publicados.

Un estudio se considera reproducible si el texto del artículo viene acompañado de los datos originales y los códigos que permiten recrear exactamente todos los resultados y figuras incluidos en el artículo. Los códigos son textos interpretables por un ordenador, y pueden ser desarrollados en R, r-*markdown*, phyton o como fichero de algún SWS (*Scientific Workflow System*) como Kepler, Taberna, Pegassus, Triana o muchos otros programas especializados (Zhao *et al.,* 2008; Talia, 2013). Por ello, un elemento de confiabilidad ante un trabajo científico sería su reproducibilidad. Es decir, que los autores dieran acceso libre, junto con la publicación, a los elementos necesarios para que los revisores y lectores pudieran repetirlo y así verificar la corrección del procedimiento o incluso probar alternativas de análisis estadísticos para asegurarse de que la conclusión es sólida y no solo un artificio del método de análisis empleado originalmente.

La reproducibilidad se relaciona directamente con la transparencia, trazabilidad y completitud del protocolo de investigación y, por tanto, es una garantía de calidad. No es una cualidad binaria, sino que existe un gradiente desde los trabajos tradicionales que son solo textos con los resultados finales (y que son totalmente irreproducibles) hasta los artículos con información suplementaria en forma de textos explicativos, métodos ampliados al detalle, los datos compartidos, los códigos del análisis y el control de las versiones, que serían los estudios perfectamente replicables (Goring *et al.,* 2013; FitzJohn *et al.,* 2014b).

La meta de promover la reproducibilidad de los estudios científicos no es luchar contra las conductas inadecuadas, sino que está pensada para la identificación de potenciales errores honestos en el procesamiento. Aun sin considerar la presencia de fraudes

deliberados o malas prácticas, la complejidad creciente de los análisis incrementa la posibilidad de errores que muchas veces no son sencillos de detectar en los textos de los trabajos o requieren de sofisticados análisis que solo pueden ser conducidos con los datos originales.

La necesidad de reproducibilidad se incrementa con el aumento del volumen de datos de los trabajos, con las nuevas tecnologías para compilar y unificar múltiples fuentes de datos complejos y altamente dimensionales, cuyo análisis con las herramientas más modernas y potentes también está llevando a mayores probabilidades de detección de asociaciones espurias. Errores e incomprensiones sobre el funcionamiento de programas científicos especializados también puede llevar a la generación de resultados incorrectos (Dominici *et al.*, 2002). La reproducibilidad no es una panacea pero incluso parcial, es mejor que nada (FitzJohn *et al.,* 2014a).

Desde el desatamiento de esta crisis, numerosos proyectos nacionales e internacionales se han desarrollado para evaluar su efecto en muchos campos. Destacan el *Reproducibility Project* en Psicología (*Open Science Collaboration*, 2015), la *Brazilian Reproducibility Initiative* en 2018 (Amaral *et al.,* 2018), iniciativas en economía experimental (Camerer *et al.*, 2016), en filosofía (Cova *et al.*, 2018) y en ciencias sociales (Camerer *et al.*, 2018). Entre todas estas se han encontrado tasas de replicabilidad entre 36% y 78%.

CONSIDERACIONES FINALES

Estas situaciones son alarmantes y han conducido, como respuesta, al desarrollo de una corriente actual que se denomina “Ciencia Abierta” que aboga por la total transparencia del trabajo investigativo para poder limpiar su imagen y aumentar la credibilidad (Laine *et al.*, 2007; Parker *et al.,* 2018; Ihle *et al*., 2017). Ella incluye el desarrollo de estrategias multinacionales guiadas por los mayores líderes de la industria científica para aumentar la reproducibilidad de las investigaciones, que está siendo cada vez más abrazada por las más importantes revistas científicas a nivel mundial.

Para combatir estas crisis algunas de las revistas líderes han dado un paso adelante, bien fomentando las publicaciones y discusiones sobre el tema (Fig. 4) como cambiando sus prácticas de revisión. En el año 2009, la revista *Biostatistics* estableció oficialmente medidas para asegurar la reproducibilidad de los artículos publicados, incluyendo este aspecto como uno a tener en cuenta en los elementos de las revisiones y adicionando la figura de un nuevo Editor Asociado enfocado en la evaluación de este factor en los manuscritos (Diggle y Zieger, 2010). El *Nature Publishing Group* anunció en mayo de 2013 que eliminaban la restricción de longitud para el acápite de materiales y métodos, que se revisarían con mayor escrutinio los análisis estadísticos con especialistas en el tema y promocionaba que los autores enviaran junto a los artículos los datos crudos para asegurar la revisión.

Requerimientos similares se han implementaron en las revistas de la Asociación Americana para el Desarrollo de la Ciencia (AAAS): *Science Translational Medicine* en 2013 y *Science* en 2014 (McNutt , 2014). La *US National Institutes of Health* (NIH) también tomó medidas para promover la reproducibilidad ya que esta crisis fue destapada, precisamente, en las investigaciones biomédicas (Collins y Tabak, 2014).

Sin embargo, aunque es una meta ideal reconocida, la Ciencia Abierta sigue chocando con barreras humanas, muchas veces por los propios elementos que producen las crisis. Los autores no comparten sus datos por miedo a que se le detecten errores que bloqueen la publicación de los trabajos (Stodden, 2011), o porque tienen la esperanza de continuar exprimiéndolos para sacar otros trabajos posteriores en el afán por publicar más, o tienen miedo a que investigadores poco éticos le tomen los datos y los empleen sin reconocer sus derechos de autor o sin recibir las ventajas de la atribución de autoría. Muchas instituciones niegan o no sacan a la luz los problemas de la baja preparación estadística entre los investigadores o de las prácticas inadecuadas de investigación, para no continuar erosionando la credibilidad pública. Los editores y revisores no desarrollan otros métodos para mejorar los filtros de calidad ante los estudios como, por ejemplo, a través de sistemas automatizados de detección de plagios, para aumentar las tasas de publicación, o disminuir el tiempo y esfuerzo por cada manuscrito.

Una respuesta alternativa que está tomando auge es el del pre-registro, es decir, la publicación *a priori* de las hipótesis y los métodos de trabajo a emplear, antes de conducir el estudio en sí mismo (Nosek *et al.,* 2018). El pre-registro de las hipótesis de trabajo fortalece la falsabilidad en las investigaciones y previene el HARKing mientras que el pre-registro de los métodos permite controlar las tasas de error de tipo 1 (Ledgerwood, 2019).

Know when your numbers are significant

The incidence of papers in cell and molecular biology that have basic statistical mistakes is alarming. Experimental biologists, their reviewers and their publishers must grasp basic statistics, urges David L. Vaux, or sloppy science will continue to grow.

(Vaux, 2012, Nature 492)

Many hands make tight work

Crowdsourcing research can balance discussions, validate findings and better inform policy, say Raphael Silberzahn and Eric L. Uhlmann.

(Silberzahn y Uhlmann, 2015, Nature 526)

Stop ignoring misconduct

Efforts to reduce irreproducibility in research must also tackle the temptation to cheat, argue Donald S. Kornfeld and Sandra L. Titus.

(Kornfeld y Titus, 2016, Nature 537)

Repeating experiments is not enough

Verifying results requires disparate lines of evidence — a technique called triangulation. Marcus R. Munafò and George Davey Smith explain.

(Munafò y Smith, 2018, Nature 553)

Five ways to fix statistics

(Nature 551)

Nature asked influential statisticians to recommend one change to improve science. The common theme? The problem is not our maths. but ourselves.

JEFF LEEK
Adjust for human cognition
Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland

MICHÈLE B. NUIJTEN
Share analysis plans and results
Tilburg University, the Netherlands

DAVID COLQUHOUN
State false-positive risk, too
University College London

STEVEN N. GOODMAN
Change norms from within
Stanford University, California

BLAKELEY B. MCSHANE AND ANDREW GELMAN
Abandon statistical significance
Northwestern University, Evanston Illinois; Columbia University, New York.

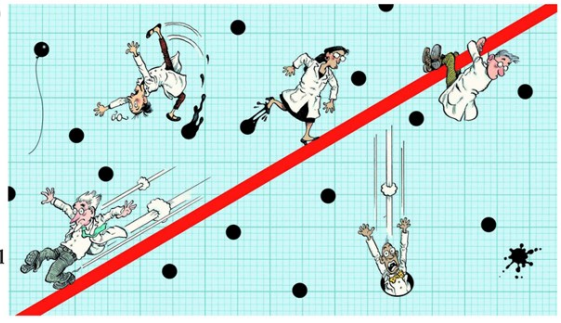

Figura 4. Muestra de algunas de las comunicaciones publicadas en *Nature* en relación a algunos de los problemas de la Ciencia.

A pesar de tener un papel fundamental en la formación ética de los futuros investigadores, la Educación Superior en Ciencias no ha tomado medidas en este sentido, aunque debe reconocerse que sí se están enfrentando con fuerza los fraudes y plagios. Button (2018) mencionaba que era chocante la revisión de las tesis de estudiantes de pregrado, a las cuales consideraba un campo de entrenamiento en malas prácticas que contribuye notablemente a diseminar entre los jóvenes investigadores estos malos hábitos. Caracterizadas por ser estudios de poco tiempo, limitados recursos, pequeños tamaños de muestra y un potencial extremo para exagerados o malos análisis estadísticos (*P-hacking*). Por la inexperiencia, los estudiantes persiguen en sus tesis el premio de la novedad, lo cual es la combinación ideal para resultados irreproducibles.

Los cursos de Metodología de la investigación muchas veces ignoran estos aspectos, y los de Ética científica son escasos (Button, 2018). Kornfeld y Titus (2016) hacían un llamado en su comunicación "*Stop ignoring misconduct*!". Byrne (2019) dijo: "*We need to talk about systematic fraud! Software that uncovers suspicious papers will do little for a community that does not confront organized research fraud*". No hablar sobre estos problemas no hará que desaparezcan y, de hecho, puede ser algo esencial para contribuir a resolverlos, ya que contribuye a aumentar la concientización y a promover el uso de mecanismos para evitarlos o controlarlos. Tenemos que hablar de los problemas, sobre todo porque es la única forma de abrir el camino a las soluciones.

Hay que tomar medidas para promover la replicabilidad y la transparencia en la Ciencia que hacemos. Hay que aprender a promover y aceptar la crítica correctiva profunda de los estudios, y verla como oportunidad de mejora. Es éticamente superior decir con la frente en alto: aquí está todo para que puedan evaluar mi trabajo en detalle, me puedo equivocar pero soy honesto en primer lugar. Como Stark (2018) dijo: la Ciencia debe ser más de "demuéstrame" que de "créeme", debe ser de "ayúdame si puedes" y no de "páreme si puedes". Y como dijo José Martí: "*La honradez debía ser como el aire y como el sol, tan natural que no se tuviera que hablar de ella*".

LITERATURA CITADA

Abbott, A., Dalton, R. y Saegusa, A. (1999). Science comes to terms with the lessons of fraud. Nature, 398(6722): 13-17.

Agnoli, F.; J.M. Wicherts, C.L.S. Veldkamp, P. Albiero y R. Cubelli. (2017). Questionable research practices among Italian research psychologists. PLoS One 12: 1–17. https://doi.org/10.1371/journal.pone. 0172792 PMID: 28296929

Amaral, O.B.; K. Neves, A.P. Wasilewska-Sampaio y C.F.D. Carneiro. (2018). The Brazilian Reproducibility Initiative. eLife 8:e41602. https://doi.org/10.7554/eLife.41602.001

Amrhein, V., S. Greenland y B. McShane. (2019). Retire statistical significance. Nature 567: 305

Arst, H.N. (2000). Apathy rewards misconduct - and everybody suffers. Nature, 403(6769): 478.

Baker, M. (2016). Is there a reproducibility crisis? Nature 533: 452-454

Barba, L. A. (2018). Terminologies for reproducible research. Preprint at https://arxiv. org/abs/1802.03311.

Broad, W. y N. Wade. (1982). Betrayer's of the truth: fraud and deceipt in the halls of science. New York: Simon and Schinter.

Button, K. (2018). Reboot undergraduate courses for reproducibility. Nature 561: 287

Button, K. S., J. P. A. Loannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S. J. Robinson y M. R. Munafò. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neur. 14: 365

Byrne, J. (2019). We need to talk about systematic fraud. Nature 566: 9

Camerer, C.F.; A. Dreber, E. Forsell, T.H. Ho, J. Huber, *et al.* (2016). Evaluating replicability of laboratory experiments in economics. Science 351:1433–1436. https://doi.org/10.1126/ science.aaf0918

Camerer, C.F.; A. Dreber, F. Holzmeister, T.H. Ho, J. Huber, *et al.* (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2: 637–644. https://doi.org/10.1038/s41562-018-0399-z

Collins F.S. y L.A. Tabak. (2014). Policy: NIH plans to enhance reproducibility. Nature 505:612–613. https://doi.org/10.1038/505612a, PMID: 24482835

Cova, F.; B. Strickland, A. Abatista, A. Allard, J. Andow, *et al.* (2018). Estimating the reproducibility of experimental philosophy. Review of Philosophy and Psychology. https://doi.org/10.1007/s13164-018-0400-9

Crimson, L. (1994). A very important erratum? - 20 years later. The New York Times, March 29, 1994.

Csada, R.D.; S. Cres, C.S. James y W. Branch. (1996). The "file drawer problem" of non-significant results: does it apply to biological research? OIKOS 76: 591–593. https://doi.org/10.2307/3546355

Davey, G. y S. Ebrahim. (2002). Data dredging, bias, or confounding. BMJ. 325 (7378): 1437–1438.

Delgado-López-Cózar, E.; D. Torres-Salinas y A. Roldán-López. (2007). El fraude en la ciencia: reflexiones a partir del caso Hwang. El profesional de la información, 16(2): 143-150. https://doi.org/10.3145/epi.2007.mar.07

Denis, D. (2017). Editorial. Hacia el rescate del concepto de autor. Rev. Cub. Cien. Biol. 6(1): 1-8

Diggle, P. J. y S. L. Zeger. (2010). Biostatistics: Editorial. Biostatistics, 11(3): 375.

Dominici, F., A. McDermott, S.L. Zeger y J.M. Samet. (2002). On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. 156: 193–203

Eriksson, S. y G. Helgesson. (2016). The false academy: predatory publishing in science and bioethics. Med. Health Care Phil. 20 (2): 163–170.

Fanelli D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics 90: 891 - 904. https://doi.org/10.1007/s11192-011-0494-7

Fidler, F.; Y. E. Chee, B. C. Wintle, M.A. Burgman, M.A. McCarthy y A. Gordon. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. Bioscience 67: 282 –289. https://doi.org/10.1093/biosci/ biw159 PMID: 28596617

Fiedler, K. y N. Schwarz. (2016). Questionable Research Practices Revisited. Soc. Psychol. Personal. Sci. 7: 45–52. https://doi.org/10.1177/1948550615612150

Fisher, R.A. (1935) The Design of Experiments. Edimburgo. Oliver and Boyd.

Fiske, S.T. y C. Dupree. (2014). Gaining trust as well as respect in communicating to motivated audiences about science topics. PNAS 111. suppl. 4: 13593–13597

FitzJohn, R.G., M.W. Pennell, A.E. Zanne, P.F. Stevens, D.C. Tank y W.K. Cornwell. (2014a). How much of the world is woody? J. Ecol 102: 1266-1272.

FitzJohn, R.G.; M. W. Pennell, A. E. Zanne y W.K. Cornwell. (2014b). Reproducible research is still a challenge. OpenScience Blog. URL http://ropensci.org/blog/2014/06/09/ reproducibility. Consultado el 17.04.2020

Forstmeier, W. y H. Schielzeth. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. Behav. Ecol. Sociobiol. 65: 47–55. https://doi.org/10.1007/s00265-010-1038-5 PMID: 21297852

Fraser, H., T. Parker, S. Nakagawa, A. Barnett y F. Fidler. (2018). Questionable research practices in ecology and evolution. PLOS ONE. https://doi.org/10.1371/journal.pone.0200303

Gelman, A. (2018). Essay: The Experiments Are Fascinating. But Nobody Can Repeat Them. Science Times at 40. The New York Times. Nov. 19, 2018

Gelman, A. y E. Loken. (2014). The Statistical Crisis in Science". Am. Sci., 102:460–465.

Gelman, A. y H. Stern. (2006). The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat. 60, 328–331

Gilbert, K.J., R.L. Andrew, D.G. Bock, M.T. Franklin, *et al.* (2012). Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. Mol. Ecol 21(20): 4925-4930.

Goring, S., T. Lacourse, M.G. Pellatt, R.W. Mathewes. (2013). Pollen assemblage richness does not reflect regional plant species richness: a cautionary tale. J. Ecol. 101: 1137-1145

Halsey, L.G., D. Curran-Everett, S.L. Vowler y G. B Drummond. (2015). The fickle P value generates irreproducible results. Nat. Methods 12 (3): 179-185

Hartemink, A.E. 2000. Publish or Perish – Fraud and ethics. Bull. Intern. U. Soil Sci. 97: 36-45

Head, M. L.; L. Holman, R. Lanfear, A.T. Kahn, M. D. Jennions. (2015). the extent and consequences of p-hacking in Science. PLOS Biology. https://doi.org/10.1371/journal.pbio.1002106

Hernández-Chavarría, F. (2007). Fraude en la autoría de artículos científicos. Rev. Biomed. 18: 127-140

Horton, R. (2015). Offl ine: What is medicine's 5 sigma? The Lancet 385: 1380

Hulbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. Ecol. Monog. 54(2): 187-21 1

Hwang, W.S., S. I. Roh, B. Chun Lee, S. K. Kang, *et al.* (2005). Patient-specific embryonic stem cells derived from human scnt blastocysts. Science 308: 1777 – 1783

Ihle, M.; I.S: Winney, A. Krystalli y M. Croucher. (2017). Striving for transparent and credible research: Practical guidelines for behavioral ecologists. Behav Ecol. 2017; 28: 348–354. https://doi.org/10.1093/beheco/ arx003 PMID: 29622916

Ince, D.C.; L. Hatton y J. Graham-Cumming. (2012). The case for open computer programs. Nature 482: 485-488

Ioannidis, J.P.A. (2005). Why most published research findings are false. PLoS Med. 2: 0696–0701. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

Ives, R. A. (2018). Informative irreproducibility and the use of experiments in Ecology. BioScience 68(10): 746-747.

John, L.K.; G. Loewenstein y D. Prelec. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 23: 524±532. https://doi.org/10.1177/0956797611430953 PMID: 22508865

Kaelin, W. G. Jr. (2017). Publish houses of brick, not mansions of straw. Nature 545: 387.

Kaiser J. (1995). Commission proposes new definition of misconduct. Science. 269:1811

Kearney, M.H. (2015). Predatory Publishing: What Authors Need to Know. Res. Nurs. Health. 38 (1): 1-3.

Kenett, R. S. y G. Shmueli. (2015). Clarifying the terminology that describes scientific reproducibility. Nat. Methods 12(8): 699

Kerr, N. (1998). HARKing: hypnothesizing after the results are known. Personal Soc Psychol Rev. 1998; 2: 196– 217. https://doi.org/10.1207/s15327957pspr0203_4 PMID: 15647155

Veld D.S. y S. L. Titus (2016). Stop ignoring misconduct. Nature 537: 29

Laine, C.; S.N. Goodman, M.E. Griswold y H.C. Sox. (2007). Reproducible research: moving toward research the public can really trust, Annals of Internal Medicine, 146: 450- 453

LeBel, E.P., D. Borsboom, R. Giner-Sorolla, F. Hasselman, *et al.* (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in Psychology. Perspect. Psychol. Sci. 8: 424–432. https://doi.org/10.1177/1745691613491437

Ledgerwood, A. (2019). The preregistration revolution needs to distinguish between predictions and analyses. PNAS Laters articles. www.pnas.org/cgi/doi/10.1073/pnas.1812592115

Leek, J. (2014). On the scalability of statistical procedures: why the p-value bashers just don't get it. Simply Statistics blog. Disponible en: http://simplystatistics.org/2014/02/14/on-thescalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/. Último acceso: 17 de abril de 2020.

Leek, T. y R. D. Peng. 2015. P values are just the tip of the iceberg. Nature, 520: 612

Lock, S., F. Wells y M. Farthing. 2001. Fraud and Misconduct in Biomedical Research. 3ra ed., BMT Books, London.

Masterson, A. (2018). At least half of evolutionary biologists and ecologists fudge results. COSMOS The Science of everything. News, Biology, 27 Marzo. Disponible en: https://cosmosmagazine.com/biology/at-least-half-of-evolutionary-biologists-and-ecologists-fudge-results-survey-finds. Último acceso: 17 de abril de 2020.

McNutt, M. (2014). Reproducibility. Science 343, 229

Mills, T. J. 2000. Position advocacy by scientists risks science credibility and may be unethical. Northw. Sci. 71(2): 165-168

Munafò, M. R. y G. D. Smith (2018). Repeating experiments is not enough. Nature 553: 399

Murphy, K. R. y H. Aguinis. (2019). HARKing: how badly can cherry-picking and question trolling produce bias in published results? J. Bus. Psych. 34(1): 1-17.

Nayernouri, T. 2009. Fraud and Dishonesty in "Scientific" Publication. Arch. Iranian Med. 12 (1): 1-4

Nosek, B.A.; C.R. Ebersole, A.C. DeHaven y D.T. Mellor. (2018). The preregistration revolution. Proc Natl Acad Sci USA 115:2600–2606.

Nuzzo, R. 2015. How scientists fool themselves–and how they can stop. Nature News, 526(7572): 182

O'Boyle, E.H., G.C. Banks y E. González-Mule. (2017). The Chrysalis Effect: How ugly initial results metamorphosize into beautiful articles. J Manage. 43: 376–399. https://doi.org/10.1177/0149206314527133

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349: aac4716. DOI: https://doi.org/10.1126/science. aac4716, PMID: 26315443

Parker, T.H. y S. Nakagawa. (2014). Mitigating the epidemic of type I error: ecology and evolution can learn from other disciplines. Front Ecol Evol. 2: 1–3. https://doi.org/10.3389/fevo.2014.00076

Parker, T.H., S.C. Griffith, J.L. Bronstein, F. Fidler, S. Foster, H. Fraser, *et al.* (2018). Empowering peer reviewers with a checklist to improve transparency. Nat. Ecol. Evol. 2.6: 929-935.

Peng, R. D. (2009). Reproducible research and biostatistics. Biostatistics 10:405–408.

Reardon, S. (2016). A mouse's house may ruin experiments. Nature News 12 – feb. https://www.nature.com/articles/nature.2016.19335

Rodríguez-Sánchez, F., A.J. Pérez-Luque, I. Bartomeus y S. Varela. (2016). Ciencia reproducible: qué, por qué, cómo? Ecosist. 25(2): 83-92

Schloss, P.D. (2018). Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. American Society for Microbiology, mBio 9(3):e00525-18

Schnitzer, S.A. y W.P. Carson. (2016). Would Ecology Fail the Repeatability Test? Bioscience 66: 98–99. https://doi.org/10.1093/biosci/biv176

Schulz, P.C. e I. Katime. (2003). Los fraudes científicos. Rev. Iberoam. Polím. 4(2): 90 pp

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., *et al.* (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives?. J. Exp. Psych. Gen. 144(6), e142.

Siegfried, T. (2014). To make science better, watch out for statistical flaws. ScienceNews. Disponible en: https://www.sciencenews.org/blog/context/make-science-better-watch-out-statistical-flaws. Ultima consulta: 4 de abril de 2020.

Simmons, J.P., L.D: Nelson y U. Simonsohn. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 22: 1359–1366. https://doi.org/10.1177/0956797611417632 PMID: 22006061

Smaldino P.E. y R. McElreath. (2016). The Natural Selection of Bad Science. https://doi.org/10.1098/rsos.160384

Stapel, D. (2012). Faking Science: A True Story of Academic Fraud. Traducción del Holandés original por Nicholas J. L. Brown (Título original "Derailment").

Stark, P. B. (2018). Before reproducibility must come preproducibility. Nature, 557(7706), 613-614.

Starr, B. (2014). Why Scientists are Seen as Competent but Untrustworthy (and Why it Matters). Science. Oct 6, 2014.

Steinschneider A. (1972). Prolonged Apnea and the Sudden Infant Death Syndrome: Clinical and Laboratory Observations. Pediatrics 50(4): 646.

Sterling, T. D., W. L. Rosenbaum y J. J. Weinkam. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. Am. Stat. 49, 108–112

Sterling, T.D. (1969). Publication decisions and their possible effects on inferences drawn from tests of significance: or vice versa. J Am Stat Assoc. 1959; 54: 30-34. https://doi.org/10.1080/01621459.1959.10501497

Stodden, V. (2011). Trust your science? open your data and code. Am. Stat News.

Talia, D. (2013). Workflow Systems for Science: Concepts and Tools. Review Article. ISRN Software Engineering. Volume 2013, Article ID 404525, 15 pp. http://dx.doi.org/10.1155/2013/404525

Teytelman, L. (2018). No more excuses for non-reproducible methods. Nature 560: 411

Thornton, A. y P. Lee. (2015). Publication bias in meta-analysis: its causes and consequences. J. Clin. Epidem. 53: 207–21

Tilburg University. (2011). Interim report regarding the breach of scientific integrity committed by prof. D.A. Stapel. Tilburg University: 1-21.

Wasserstein, R.L. y N. A. Lazar. (2016). The ASA's statement on p -values: context, process, and purpose, The American Statistician, Vol-Pag http://dx.doi.org/10.1080/00031305.2016.1154108

Wasserstein, R.L., A.L. Schirm y N.A. Lazar. (2019). Moving to a World Beyond "p<0.05". The American Statistician, 73(sup1): 1 -19, http://dx.doi.org/10.1080/00031305.2019.1583913

Weltzin, J. F., R.T. Belote, L. T. Williams, J. K. Keller y E. C. Engel. (2006). Authorship in ecology: attribution, accountability, and responsibility. Front Ecol Environ 4(8): 435–441

Wooster, W. S. (1998). Science, Advocacy, and Credibility. Science 282: 1823 - 1824

Zhao, Y., L. Raicu e I. Foster. (2008). Scientific Workflow Systems for 21st Century, New Bottle or New Wine? Invited Short Paper. http://dx.doi.org/10.1109/SERVICES-1.2008.79

• • •