Traducción al castellano de Blanco, F., Perales, J.C., & Vadillo,M.A. (2017). Pot la psicologia rescatar-se a si mateixa? Incentius, biaix i replicabilitat. *Anuari de psicologia de la Societat Valenciana de Psicologia*, 18 (2), 231-252.
http://roderic.uv.es/handle/10550/21652 DOI: 10.7203/anuari.psicologia.18.2.231

¿PUEDE LA PSICOLOGÍA RESCATARSE A SÍ MISMA? INCENTIVOS, SESGOS Y REPLICABILIDAD

CAN PSYCHOLOGY RESCUE ITSELF? INCENTIVES, BIASES, AND REPRODUCIBILITY

Fernando Blanco[1,4], José C. Perales[2,4], Miguel A. Vadillo[3,4]

1. Departamento de Fundamentos y Métodos de la Psicología, Facultad de Psicología y Educación, Universidad de Deusto. fernandoblanco@deusto.es

2. Departamento de Psicología Experimental, Centro de Investigación Mente, Cerebro y Comportamiento (CIMCYC), Universidad de Granada. jcesar@ugr.es

3. Departamento de Psicología Básica, Universidad Autónoma de Madrid. miguel.vadillo@uam.es

4. Sociedad para el Estudio de los Juicios y las Decisiones (SEJyD), @SociedadEJyD

Autor para correspondencia:

José C. Perales
Departamento de Psicología Experimental
Universidad de Granada.
Campus de Cartuja, s/n
18071, Granada, España
(+34) 958 247882
jcesar@ugr.es

Traducción al castellano de Blanco, F., Perales, J.C., & Vadillo,M.A. (2017). Pot la psicologia rescatar-se a si mateixa? Incentius, biaix i replicabilitat. *Anuari de psicologia de la Societat Valenciana de Psicologia*, 18 (2), 231-252.
http://roderic.uv.es/handle/10550/21652 DOI: 10.7203/anuari.psicologia.18.2.231

Resumen

En los últimos años la ciencia psicológica está sufriendo, interna y externamente, una importante crisis de credibilidad, a la que tampoco han sido ajenas otras ciencias como la Medicina o la Biología. Varios proyectos de ciencia colaborativa sugieren que gran parte de los resultados de la investigación en la disciplina son difíciles de reproducir. Esto viene acompañado de un conjunto de simulaciones y análisis cuantitativos que sitúan la cantidad falsos positivos por encima del 50% del total de datos publicados en la literatura psicológica actual.

En este breve ensayo realizamos un análisis actualizado de esta situación, e intentamos identificar las causas psicológicas que han contribuido a la misma. Entre esas causas destacan, primero, un sistema de incentivos alineado con los intereses de promoción del investigador y no con el descubrimiento y diseminación de ciencia transparente, fiable y reproducible y, segundo, los sesgos individuales en la elaboración de juicios y toma de decisiones que afectan de forma generalizada a los seres humanos (sean o no investigadores).

Terminamos discutiendo cómo estos sesgos, en concordancia con la investigación clásica sobre sesgos y heurísticos, se acumulan a través de individuos y acaban afectando de forma sustancial a la empresa colectiva de la ciencia psicológica. Enumeramos también las estrategias que podrían contribuir, de implementarse con éxito, a corregir en parte la situación, e incluso extenderse a otras áreas científicas.

Palabras clave: Replicabilidad, Sesgos, Métodos en Psicología, Incentivos, Ciencia abierta

Traducción al castellano de Blanco, F., Perales, J.C., & Vadillo,M.A. (2017). Pot la psicologia rescatar-se a si mateixa? Incentius, biaix i replicabilitat. *Anuari de psicologia de la Societat Valenciana de Psicologia*, 18 (2), 231-252.
http://roderic.uv.es/handle/10550/21652 DOI: 10.7203/anuari.psicologia.18.2.231

Abstract

In the last years, the credibility of psychological science – as well as the one of other disciplines as biology or medicine – has been called into question. Results from a number of collaborative science projects suggest that a large portion of results in published psychological research is not reproducible. In parallel, simulation and quantitative inference methods estimate the proportion of false positives in current psychological literature to be well above 50%.

In this essay we present an up-to-date analysis of the current situation and its psychological underpinnings. Among these psychological factors, we highlight the importance of (1) a career promotion-centred incentive system that disregards scientific transparency and dissemination of reliable and reproducible research, but strongly motivates researchers to strive for publication, and (2) individual cognitive biases and distortions – common to laypeople and scientists – that affect research-related judgment and decision making.

We conclude discussing how these cognitive distortions, in line with available evidence on heuristics and biases, accumulate across researchers and end up distorting evidence accrual, and thus making the enterprise of collective psychological science derail. Awareness of these consequences and their roots should help implement strategies designed to correct the current distortions and even make psychological science a role model for other disciplines experiencing similar problems.

Key words: Reproducibility, Biases, Psychology methods, Incentives, Open Science

Introducción

En un artículo de la revista Nature, Sir Francis Galton (1907a) relata un famoso estudio de campo en el que recogió las estimaciones realizadas, sin ningún instrumento de medida, por 787 personas sobre el peso de un buey. Entre esas personas se encontraban ganaderos y carniceros, pero también un número considerable de ciudadanos de a pie sin experiencia en tales cuestiones. La estimación media fue de 1207 libras, una cifra muy cercana a las 1198 libras que resultó finalmente pesar el buey.

El mismo Galton (1907b) utilizó esta propiedad de los juicios colectivos para defender los principios de la democracia y posteriormente, se ha utilizado para defender la naturaleza acumulativa y autocorrectiva de la Ciencia. Por ejemplo, Surowiecki (2004) relata sobre la búsqueda de la causa de la enfermedad respiratoria SARS:

> "A principios de Febrero de 2003 […] a varios laboratorios se les encargó trabajar al mismo tiempo sobre las mismas muestras, multiplicando su velocidad y su eficacia. Tras unos días de esfuerzo, los laboratorios consideraron y descartaron un cierto número de causas […]. El 21 de Marzo de 2003, científicos de la Universidad de Hong Kong habían aislado un candidato. […] Durante la semana siguiente, los laboratorios de la red detectaron el coronavirus en una amplia variedad de muestras" (p. 159).

En otras palabras, el principio de sabiduría de las masas predice que cuando varios agentes independientes aportan información a un conjunto, aquello que los agentes aportan de cierto tiende a acumularse, mientras que los errores tienden a eliminarse. Esta eliminación del error por promediado opera en muchos contextos distintos y explica, por ejemplo, la eficiencia de los mercados de decisión (a los que volveremos al final del este artículo). En el ámbito que nos incumbe, ese mismo principio está en la base de la principal fuente de información necesaria para construir una Psicología basada en la evidencia –los meta-análisis y las revisiones sistemáticas–. Estos trabajos acumulativos se basan en estudios individuales realizados por investigadores y laboratorios en todo el mundo (Higgins y Green, 2008; Cumming, 2014).

Sin embargo, a menudo se olvida que la correcta operación de ese principio requiere que se cumplan varias condiciones. La primera es que el muestreo de los agentes individuales y de la información que aportan no sea selectivo. La segunda, que los errores que cometen cada uno de esos actores sean aleatorios (no tiendan a estar sistemáticamente desviados en la misma dirección). Y la tercera, que los actores y sus juicios sean independientes entre ellos (Lorenz, Rauhut, Schweitzer y Heldbing, 2011). Por desgracia, ninguna de esas tres condiciones se cumple actualmente ni en la Psicología científica ni en la Ciencia en general, por lo que podemos sospechar que ambas están en gran medida fracasando como empresas acumulativas y autocorrectivas (Ioannidis, 2005, 2012). Este breve ensayo pretende ilustrar cómo factores psicológicos y ambientales que afectan de forma concreta a las decisiones de investigadores individuales son en última instancia responsables de esa situación. Además, apuntaremos algunos elementos que podrían contribuir a corregirla si se implementan de forma eficaz. Esto es, pretendemos que la Psicología se embarque en la misión de describir, explicar e intervenir en la corrección de sus propios sesgos y puntos ciegos.

La crisis de la Psicología científica y sus causas inmediatas

En buena medida, la credibilidad de una ciencia depende de la posibilidad de reproducir los efectos reportados en sus estudios. Hablamos de replicación metodológica al éxito en reproducir un efecto exactamente en las mismas condiciones en las que se describió el efecto original. La replicación conceptual, por el contrario, supone la reproducción en condiciones distintas a las originales (en otra población, otra tarea u otro contexto, por ejemplo), pero en las que se supone que opera el mismo proceso psicológico subyacente. Un fracaso en la replicación conceptual limita el alcance teórico de un resultado y, por tanto, también su posible utilidad práctica, pero no tiene por qué afectar a la veracidad del resultado original. Un fracaso en una réplica metodológica tampoco es inherentemente dañino; puede indicar que el resultado original o el del intento de réplica se debió al azar (una posibilidad que es inherente al establecimiento de umbrales de significatividad estadística). Sin embargo, un fracaso sistemático en la replicación metodológica puede ser fatal para la credibilidad de una disciplina, en tanto que puede indicar una tendencia también sistemática en los resultados originales a ser falsos.

En los últimos años, varios intentos por reproducir efectos que eran centrales en determinados modelos teóricos han fracasado. Entre esos efectos están por ejemplo el de la pose de poder y otros efectos similares de corporización (Ranehill, Dreber, Johannesson, Leiberg, Sul y Weber, 2015), los de depleción del ego (Lurquin et al., 2016), muchos de los efectos de facilitación social (e.g. Rohrer, Pashler y Harris, 2015; Shanks et al., 2015) y una buena parte de las intervenciones psicoterapéuticas en distintos contextos de aplicación (e.g. Van der Gucht et al., 2016; Coyne, Thombs y Hagedoorn, 2010).

El problema de la fiabilidad de la literatura psicológica no es nuevo y ha sido señalado en muchas ocasiones, de forma más o menos aislada, a lo largo de las últimas décadas (e.g. Meehl, 1978). La comunidad científica en su conjunto, sin embargo, ha empezado a tomar conciencia del problema en los últimos años. Por ejemplo, en 2012, Daniel Kahneman –premio Nobel y probablemente el psicólogo vivo más influyente en la actualidad– envió una carta abierta en la que advertía de la fragilidad de prácticamente la totalidad de la investigación sobre facilitación social y advertía a los jóvenes científicos de los problemas para su futura carrera que podía conllevar investigar en esa área (Young, 2012).

Ante estos problemas en ocasiones se ha argumentado, primero, que afectaban sólo a ciertos tópicos de investigación, como la facilitación social y la literatura circundante en juicios y toma de decisiones. Y, segundo, que sobre todo se trataba de un fracaso en la réplica conceptual y que, por tanto, no necesariamente afectaba a la veracidad de los descubrimientos originales. El intento más serio, a gran escala, de cuantificar los problemas de replicabilidad metodológica en el conjunto de la Psicología ha sido fruto de la colaboración entre múltiples investigadores y laboratorios, bajo el paraguas de la Open Science Collaboration (2015). En ese proyecto se intentaron 100 réplicas de estudios extraídos aleatoriamente de tres de las revistas más prestigiosas, utilizando diseños de alta potencia y contando con la colaboración de los autores de los estudios originales en el procedimiento y la elaboración de materiales. De los 100 estudios originales, 97 reportaban resultados estadísticamente significativos ($p<0.05$) y sin embargo, en sólo un 39% de ellos el intento de réplica alcanzó ese nivel de significación estadística (esto es, no se replicó el efecto). Y lo que es más grave, el

tamaño medio del efecto observado en las réplicas fue menos de la mitad del observado en los estudios originales, y en sólo un 47% de los casos el tamaño observado del efecto estuvo dentro del intervalo de confianza (95% IC) estimado en el estudio original. Decimos que es especialmente grave porque el tamaño del efecto y no la p, a pesar de lo que muchos científicos piensan, es el resultado más relevante de un estudio y el único a partir del cual se puede valorar la relevancia teórica y práctica de ese efecto y acumular evidencia a través de estudios individuales (Cohen, 1990).

A grandes rasgos, estos resultados coinciden con otros proyectos colaborativos (e.g. Klein et al., 2014). Aunque se han interpretado de forma más o menos optimista por distintos autores (Baker, 2015; Van Bavel, Mende-Siedlecki, Bradi y Reinero, 2016), parece difícil negar que una porción muy importante de los resultados en Psicología son falsos positivos. La razón inmediata de ello resulta evidente en las cifras aportadas arriba: un 97% de los efectos originales de la muestra de la Open Science Collaboration se basaban en diferencias estadísticamente significativas (positivos). Ello a pesar de que la potencia media de los estudios en Psicología no alcanza el 50% (Bakker, Van Dijk y Wicherts, 2012; Button et al, 2013; Cohen, 1990). Ello implica que más del 50% de los estudios en los que se intenta demostrar una hipótesis alternativa verdadera no tienen éxito y nunca salen a la luz. Este ocultamiento de los resultados nulos (Franco, Malhotra y Simonovits, 2014; Sterling, Rosenbaum y Weikram, 1995) en ocasiones afecta a estudios completos. Otras veces afecta a porciones de estudios en los que se utilizan varias variables, de las que sólo se reconocen y reportan aquellas que dan lugar a diferencias significativas. Por último, ese ocultamiento puede referirse a estrategias de análisis de entre un menú disponible, de las cuales sólo se reconoce haber llevado a cabo una.

En este contexto, interesa saber cuántos de los resultados reportados son resultados verdaderos. El porcentaje de verdaderos positivos sobre el total de efectos estadísticamente significativos reportados en la literatura de un determinado tópico recibe el nombre de valor predictivo positivo (PPV, su acrónimo en inglés). Según Ioannidis (2005), es posible estimar el PPV en un campo determinado simulando adecuadamente las prácticas de investigación en ese campo y algunas de sus características inherentes. Entre ellas están (1) la potencia habitual de los estudios (que

depende principalmente del tamaño de las muestras y del tamaño del efecto que se persigue), (2) la novedad y valor contraintuitivo de los efectos que se pretenden demostrar, (3) la flexibilidad de posibilidades de análisis, y (4) la excesiva libertad en la definición y selección de variables dependientes.

Así, los ensayos clínicos de alta potencia (1-β=0.80), con variables de resultado y análisis pre-especificados y con hipótesis bien informadas podrían alcanzar un PPV de 0.85, mientras que, en el otro extremo, la gran masa de análisis exploratorios, con muestras pequeñas (1-β=0.25), sobre efectos improbables a priori y análisis selectivos estaría cercana a un PPV de 0. Desgraciadamente, también es fácil demostrar que la mayor parte de los trabajos que se publican en la actualidad están más cerca del segundo extremo que del primero (el lector puede fácilmente simular el comportamiento de distintos escenarios en la aplicación online desarrollada por Zehetleitner y Schönbrodt, 2016). Lo que ello implica es que, en la mayor parte de los campos de estudio, los meta-análisis y las revisiones sistemáticas que deberían servir para basar la práctica en la evidencia son virtualmente irrelevantes, porque aglutinan datos estadísticos profundamente sesgados.

Causas intermedias: mecanismos psicológicos responsables de las prácticas de investigación cuestionables

Conocimiento estadístico

Aunque la causa más importante de la baja fiabilidad de la investigación psicológica es la publicación casi exclusiva de resultados positivos obtenidos en condiciones subóptimas y propensas a la distorsión sistemática, las razones subyacentes tienen que ver con nuestros comportamientos y decisiones como investigadores. Voluntaria o incidentalmente, empujamos de forma sistemática los resultados de nuestra investigación en la dirección que mejor sirve a nuestros intereses.

Uno de los mecanismos que perpetúan estos problemas es que los propios científicos carecemos de la formación o del criterio necesario para juzgar la calidad de nuestros métodos. Por ejemplo, como hemos señalado con anterioridad, la potencia estadística es un factor determinante del valor predictivo positivo de toda investigación.

Los estudios que se han realizado hasta la fecha sugieren que la potencia estadística de la investigación psicológica es ridículamente baja (Bakker et al., 2012; Button et al, 2013; Cohen, 1990). Esta situación no sólo no parece mejorar con el paso de los años, sino que de hecho tiende a empeorar con cada nueva evaluación. Como ya señalaban Sedlmeier y Gigerenzer (1989) hace casi tres décadas, los estudios sobre la potencia estadística tienen poco o ningún efecto sobre la potencia estadística de los estudios.

¿Por qué los científicos insistimos en realizar estudios sin suficiente potencia? ¿Cómo es posible que a un científico le resulte rentable realizar un estudio que sólo tiene un 30% o un 40% de probabilidades de arrojar resultados significativos, suponiendo que la hipótesis que lo motiva sea cierta? Aparentemente, los investigadores no se plantean la cuestión en estos términos. Más bien, tendemos a sobreestimar la potencia estadística que se alcanza con los tamaños muestrales y los efectos más habituales en nuestra disciplina (Bakker, Hartgerink, Wicherts y van der Maas, 2016).

La incapacidad para comprender la información estadística esencial no se limita a la potencia. El propio concepto de significación de la hipótesis nula es objeto de todo tipo de malentendidos, tal y como revela el hecho de que sólo una minoría de investigadores conoce la definición correcta del valor p (e.g. Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos y Longobardi, 2016; Badenes-Ribera, Frias-Navarro, y Pascual-Soler, 2015; Schervish, 1996). Todo sugiere que algunos aspectos cruciales de la investigación, como la selección del tamaño muestral o del procedimiento estadístico a utilizar, no se basan en ningún proceso analítico preciso, sino en el simple hábito y en intuiciones vagas que no siempre se corresponden con la realidad. En la mayor parte de los casos, es probable que tanto estas decisiones como las interpretaciones de los resultados publicados por otros investigadores se basen en heurísticos sencillos como "incluir al menos 25 participantes por condición" o "un resultado estadísticamente no significativo indica que no hay diferencias entre condiciones" (Hoekstra, Finch, Kiers y Johnson, 2006; Nelson, Rosenthal y Rosnow, 1986). Como sucede con cualquier otro heurístico, estas reglas de decisión pueden funcionar correctamente en muchas ocasiones, pero su utilización mecánica e inconsciente puede tener importantes consecuencias negativas (Gigerenzer, 2004).

Sesgos individuales en la toma de decisiones

Con todo, sería un error atribuir íntegramente la actual crisis de replicabilidad a la simple ignorancia estadística. Al fin y al cabo, la mayor parte de los científicos tenemos que lidiar en nuestro quehacer cotidiano con conceptos y herramientas considerablemente más complejos que los estadísticos. Por el contrario, cabe imaginar que buena parte de estos problemas tienen lugar porque los investigadores no somos testigos neutrales de la evidencia que recogemos, sino observadores interesados que percibimos nuestros propios datos a través de los mismos sesgos cognitivos que pueden encontrarse en cualquier ciudadano de a pie. Como cualquier otra persona, los científicos somos incapaces de evitar ver patrones donde sólo hay ruido y azar. Aunque los métodos de investigación están diseñados como contramedidas para anular el efecto de estos sesgos, este objetivo raramente se alcanza.

La estadística inferencial basada en la significación de la hipótesis nula requiere que el experimentador defina a priori qué hipótesis desea poner a prueba, antes de recoger cualquier dato. Sin embargo, una vez que finaliza el experimento y se procede a analizar los datos, el investigador inevitablemente comenzará a ver patrones en ellos, algunos directamente relacionados con su hipótesis inicial y otros totalmente ajenos a ella (Gelman y Loken, 2014). Por ejemplo, un simple vistazo a los estadísticos descriptivos puede sugerir que la hipótesis inicial se cumple, pero sólo para una parte de la muestra. Tal vez sólo para las mujeres, o sólo para los participantes más jóvenes, o tal vez para aquellos que puntúan más alto en neuroticismo. Posiblemente ninguno de estos factores formaba parte del plan de análisis que el investigador se había propuesto hacer a priori. Pero una vez vistos los estadísticos descriptivos es difícil no pensar que puedan ser relevantes. En otras palabras, con frecuencia las hipótesis experimentales no están completamente definidas antes de que se recojan los datos, sino que se van formando en la mente del investigador a medida que se exploran los datos.

Si en el proceso de explorar los datos el investigador se topa con un resultado significativo, aunque no formara parte del plan de análisis original, no le será difícil tergiversar el curso real de los acontecimientos y convencerse a sí mismo de que, en realidad, esa hipótesis se había contemplado desde el primer momento (práctica

conocida como HARKing, Hypothesis After Results are Known; Kerr, 1998). La extensa literatura sobre el sesgo de retrospección muestra que, una vez que conocemos el resultado de un evento, tendemos a pensar que ese resultado era perfectamente predecible de antemano (Arkes, 2013; Fischhoff, 1975). Es decir, una vez que sabemos algo, nos cuesta retrotraernos a la situación en la que estábamos cuando aún no lo sabíamos. Es improbable que el investigador que por accidente se topa con un resultado inesperado sea inmune a este sesgo. Seguramente, una vez vistos los datos, pensaremos que esos resultados eran predecibles y que, de alguna forma vaga, formaban parte de la hipótesis inicial.

El hecho de que las hipótesis vayan emergiendo a medida que se exploran los datos, en lugar de ser claramente formuladas a priori, invita a analizar los datos de tantas formas como sea necesario hasta que se encuentre un resultado positivo (práctica conocida como p-hacking; Wicherts et al., 2016). Las encuestas que se han realizado hasta el momento confirman que muchos investigadores maquillan los datos con prácticas como, por ejemplo, registrar varias variables dependientes pero informar sólo de las que arrojan resultados significativos (i.e. cherry picking), eliminar cualquier alusión a condiciones experimentales donde no se encontró ningún resultado significativo, o añadir covariables a un análisis estadístico con el único fin de hacer que una tendencia no significativa alcance la significación estadística.

En la literatura, estas prácticas se han dado a conocer como prácticas de investigación cuestionables (e.g. John, Loewenstein y Prelec, 2012) y tienen como consecuencia el incremento en la tasa de falsos positivos. Al combinar varias de estas prácticas es fácil alcanzar tasas de falsos positivos superiores al 60% (Simmons, Nelson y Simonsohn, 2011). Por supuesto, cualquiera de estas estrategias constituye mala praxis. Sin embargo, sólo lo es porque en todos los casos supone recurrir a un plan de análisis que se aleja de la hipótesis inicial del estudio. Si el investigador es capaz de convencerse a sí mismo de que estas estrategias de análisis se habían contemplado de antemano, entonces es perfectamente factible llevarlas a cabo sin tener en ningún momento la sensación de haber actuado mal. El sesgo de retrospección puede jugar un papel clave en este proceso.

Al sesgo de retrospección se añaden al menos otros dos sesgos cognitivos que llevan a cualquier persona a percibir la realidad de la forma que mejor se acomode a sus creencias: los sesgos de confirmación y de distorsión de la información (DeKay, 2015; Nickerson, 1998). Si un investigador ha realizado dos experimentos, de los cuales uno apoya su hipótesis y otro no, seguramente dará más peso a cualquier evidencia que sugiera que el primer resultado es más sólido que el segundo. Si el estudio con el resultado favorable tiene más participantes, se achacará la diferencia a la mayor potencia estadística. Si el estudio con el resultado negativo se realizó a una hora muy temprana, tal vez se asuma que los participantes no estaban totalmente despiertos al venir al laboratorio. Cualquier persona inteligente –y los científicos lo somos- será capaz de encontrar razones que expliquen que el resultado esperado no se haya cumplido en todas las ocasiones. En este sentido, es interesante recordar que el sesgo de publicación al que hemos aludido más arriba no se debe tanto al hecho de que las revistas no publiquen resultados negativos como al hecho de que los propios investigadores, en primera instancia, son reacios a incluir estos estudios en sus manuscritos (Dickersin, 2005). Esto podría estar revelando que los científicos dan más peso a sus experimentos cuando arrojan resultados significativos, aunque lógicamente es posible que otros factores estén contribuyendo a esta decisión.

Como comentamos en la introducción, en un mundo ideal podría esperarse que, aunque estos sesgos puedan afectar al criterio de uno u otro investigador tomados de forma aislada, colectivamente se cancelarían entre sí, de modo que la comunidad científica en su conjunto sería inmune a ellos. Por ejemplo, si el científico A es partidario de una teoría y el científico B es detractor de la misma, cabría esperar que los sesgos cognitivos de A y de B se anularan mutuamente, de tal forma que el cuerpo de evidencia recogido por ambos condujera a una visión no sesgada de la realidad. Lamentablemente, décadas de investigación en toma de decisiones grupales muestran que el consenso final al que llega un grupo rara vez se corresponde con la creencia media de sus individuos, sino que tiende a polarizarse hacia el extremo dominante (Myers y Lamm, 1976). En el contexto de la ciencia, esto sugiere que una vez que una teoría se convierte en dominante dentro de una comunidad de investigadores, aunque sea sobre la base de una evidencia débil y poco fiable, es probable que cualquier dato

discordante se reinterprete o reevalúe hasta que encaje con la teoría (tal vez con la ayuda de hipótesis auxiliares) o bien se desestime como irrelevante (Chinn y Brewer, 1998). Una evidencia indirecta de este hecho puede encontrarse en los análisis que revelan que la muerte o retiro de un investigador de referencia en un campo de investigación suele ir seguida de un crecimiento de los intentos por poner a prueba hipótesis más heréticas y menos complacientes con la opinión de la mayoría (Fons-Rosen y Azoulay, 2015).

Causas distales: el sistema de recompensas y castigos de la Psicología académica

Entre todas las características que definen la práctica actual de la ciencia, una de las más relevantes es su carácter competitivo. Dado que los recursos con los que trabajan los científicos son limitados, con frecuencia se asume que la competición basada en méritos es una manera adecuada de asignar dichos recursos (ya sean los fondos de un proyecto de investigación, la estabilidad de una plaza, el acceso a instrumental o instalaciones, o el espacio en una revista). Sin embargo, el nivel de competición que impregna la práctica científica también puede acabar implantando incentivos perversos en el sistema e instalando dinámicas no deseables, con consecuencias negativas para los investigadores y para la propia empresa del avance científico.

Como toda competición, la carrera científica actual tiene sus propias reglas para premiar a los ganadores. En este contexto, se han descrito tres indicadores que determinan el éxito profesional de un investigador: el número de publicaciones, el índice de impacto de las revistas donde se ha publicado su investigación y el número de citas recibidas (van Dijk, Manor, y Carey, 2014). Prácticamente todos los índices de calidad y productividad empleados en la gestión de la investigación se basan en estos pilares (ver Hirsch, 2005). A partir de aquí, es fácil intuir cómo el uso generalizado de estas métricas puede introducir distorsiones indeseadas en el comportamiento de los científicos. Si se plantea la ciencia como una competición no es extraño que los investigadores acaben seleccionando las mejores estrategias para ganar (Bakker et al., 2012).

En primer lugar, recompensar a un investigador por el número de publicaciones que consigue provoca que los equipos de investigación opten por escribir artículos más breves y menos elaborados, en vez de trabajos completos (con múltiples experimentos, mejores controles y muestras grandes) que aporten evidencias más concluyentes. En ocasiones, la presión por publicar conduce a la práctica de la publicación salami (Smolčić, 2013), que consiste en la descomposición de un estudio amplio en partes parcialmente redundantes (misma muestra, mismas hipótesis, y similar metodología) y su publicación separada en múltiples revistas, con el objetivo de inflar el cálculo total de trabajos publicados. Como consecuencia, estos resultados crean la falsa impresión de apoyarse unos a otros e inundan la literatura científica con un caudal muy abultado de evidencia poco relevante, dificultando sacar conclusiones claras de su revisión.

En cuanto al índice de impacto y al conteo de citas recibidas por artículo, ambos números se toman como una aproximación indirecta a la calidad del artículo, ya que la valoración directa de la misma es más compleja. Estos índices han sido frecuentemente criticados (Alberts, 2013; Brembs, Buttn y Munafò, 2013; Kirschner, 2013). El índice de impacto de una revista no siempre correlaciona con la repercusión de cada artículo individual, e incluso hay evidencia de que las investigaciones publicadas en las mejores revistas podrían ser menos fiables (por ejemplo, las revistas de mayor impacto también tienen una tasa mayor de retracciones y de errores en los valores p, Brembs et al., 2013; Szucs y Ioannidis, 2016). Pero además el uso de estas métricas como indicio de calidad de la investigación introduce artefactos en la práctica diaria del científico y crea incentivos para la manipulación misma de esos índices.

En primer lugar, dado que las revistas compiten entre sí por obtener un puesto aventajado en los rankings, su incentivo es a menudo priorizar la publicación de aquellos resultados en los que se espere una repercusión grande, fuera incluso del círculo de la prensa académica. Muchos de los efectos protagonistas en la reciente crisis de replicación de la psicología (por ejemplo, que la exposición a palabras como "jubilación" o "viejo" ralentiza la velocidad a la que el participante camina para abandonar el laboratorio, Bargh, Chen y Burrows, 1996) tienen ese ingrediente de resultado sorprendente o contraintuitivo al gusto del público y los editores de las revistas. Esta preferencia por los resultados más vendibles instala otro sesgo de

publicación de facto: no sólo los resultados significativos, sino también los que tienen más potencial mediático, se publican con mayor facilidad. Así, áreas con mucho atractivo como la neurociencia adquieren una ventaja en cuanto a número de investigadores y asignación de fondos, mientras que otras áreas o enfoques menos interesantes para el público se van poco a poco despoblando. Al mismo tiempo, se introduce un sesgo que penaliza la creatividad y el riesgo: los caminos más transitados en la ciencia actual son aquellos que conducen a la publicación en revistas de alto impacto y pocos investigadores jóvenes arriesgarían su carrera hoy para tomar otras rutas más originales.

Y en segundo lugar, este sistema de incentivos basado en métricas imperfectas es también un caldo de cultivo idóneo para la aparición de las prácticas científicas cuestionables. Una investigación sobre los efectos de la competición sobre la conducta de investigadores jóvenes (Anderson, Ronning, De Vries y Martinson, 2007) reveló que la presión por destacar en las métricas los empujaba a las siguientes prácticas problemáticas: (1) Publicación estratégica de resultados, que incluye aspectos como la descomposición de una única línea de investigación en diversos artículos, parcialmente redundantes entre sí, o la publicación selectiva (únicamente de los mejores resultados, o los significativos). (2) Tendencia a ocultar, o a no compartir, información relevante sobre las investigaciones. Por ejemplo, muchos investigadores evitan anticipar sus ideas o futuras publicaciones en congresos y reuniones científicas, o incluso recelan de los revisores que examinan sus trabajos previamente a la publicación. También se suelen ocultar detalles del procedimiento para evitar perder la ventaja ante los competidores. (3) Introducción de sesgos en el proceso de revisión por pares, en ocasiones para dificultar la tarea a los competidores. (4) Uso descuidado o poco meticuloso de las técnicas estadísticas, con el objetivo de acortar plazos y exagerar la importancia de los resultados. Como caso extremo, la presión por sostener un buen ritmo de publicaciones prestigiosas como condición necesaria para conseguir y mantener contratos, financiación y recursos puede incluso tentar a una minoría de los investigadores a cometer fraude o falsificación de datos (Zatonski y Witkowski, 2015), como se hizo patente en el célebre caso del psicólogo Diederick Stapel (Enserink, 2012). Por los motivos mencionados y otros, algunas asociaciones de científicos como la Sociedad

Americana de Biología Celular han propuesto que las métricas basadas en cálculos de impacto y citas dejen de emplearse como criterio para contrataciones, promociones y obtención de financiación en la carrera científica (American Society for Cell Biology, 2013).

Simulando el sistema de incentivos en Ciencia

Más allá de que la presión por obtener buenas puntuaciones en las métricas pueda motivar prácticas cuestionables, como hemos visto, el sistema de competición en general crea incentivos para producir estudios de baja calidad y con resultados poco fiables. Por ejemplo, Bakker et al. (2012) realizaron simulaciones por ordenador en las que compararon la ejecución de diversos agentes artificiales (que en este caso hacían el papel de investigadores) con diferentes estrategias. Por ejemplo, una posible estrategia es realizar un estudio a gran escala, con una muestra lo suficientemente grande como para alcanzar una potencia estadística aceptable. Por otro lado, otras estrategias combinan la realización de múltiples estudios más pequeños (i.e., con menor potencia) con el uso de determinadas prácticas cuestionables, como la publicación selectiva (sólo los resultados significativos se escriben). Las simulaciones mostraron cómo este segundo tipo de estrategia es mucho más eficiente a la hora de conseguir el máximo de resultados significativos en el menor tiempo.

Más recientemente, Higginson y Munafò (2016) usaron una aproximación similar, pero más completa en su recreación del sistema de incentivos, encontrando que la estrategia más racional para maximizar el número de publicaciones consiste en concentrar los esfuerzos en hallar resultados novedosos (lo cual deja fuera cualquier trabajo de replicación de estudios previos, así como estudios puramente confirmatorios), a base de estudios con muy poca potencia (entre 0.1 y 0.4). Como se mencionó previamente, los estudios con baja potencia son también los que más frecuentemente producirán resultados erróneos (sea inflando el tamaño de efectos reales, o bien produciendo falsos positivos). El comportamiento de este agente perfectamente racional, en las simulaciones, tal vez no se distancia mucho del de gran parte de los investigadores reales, cuya práctica diaria está condicionada por un sistema de

recompensas y castigos similar, a tenor de algunas investigaciones realizadas (Fanelli, 2010).

Un nuevo sistema de incentivos para una mejor ciencia

Si en la sección previa hemos argumentado que el sistema de incentivos en ciencia puede conducir a la proliferación de prácticas cuestionables, estudios sin potencia estadística y conclusiones erróneas, a continuación mencionaremos algunas de las reacciones que han tenido lugar para corregir los problemas detectados.

Políticas editoriales diferentes

Las revistas científicas son en gran medida responsables de la situación actual de la ciencia psicológica. Por ese motivo, el sistema de investigación debe ser sensible a los cambios en las políticas editoriales de las revistas. En este sentido, estamos asistiendo a algunos movimientos cuyo objetivo es mejorar la situación descrita en secciones anteriores. En primer lugar, algunas revistas ya contemplan explícitamente la posibilidad de publicar réplicas de otros estudios, o resultados nulos (ver PLoS Blogs, 2015). Unas pocas han abogado expresamente por dejar fuera la novedad, relevancia, o el impacto previsto como criterios para aceptar un manuscrito y afirman basarse únicamente en la calidad del aspecto metodológico. Sin ir tan lejos, muchas de las revistas tradicionales en psicología, como Psychological Science (APS), están ampliando el espacio dedicado a describir los métodos y procedimientos, de forma que estos puedan ser conocidos con detalle, al tiempo que fomentan nuevas perspectivas estadísticas que eviten (o al menos compensen en cierta medida) los problemas asociados a los valores p y en general a los tests de significación basados en la hipótesis nula. Entre estas nuevas perspectivas está el adecuado uso de análisis de potencia, el estudio de los tamaños del efecto con intervalos de confianza asociados o, últimamente, la estadística Bayesiana (Cumming, 2014; Etz, Gronau, Dablander, Edelsbrunner y Baribault, 2016; Wagenmakers, 2007).

Premiar por compartir

Más allá de ofrecer la posibilidad de paliar algunos problemas con el actual sistema y de desincentivar algunas malas prácticas, algunas propuestas se están centrando en fomentar, mediante incentivos, las buenas prácticas en el ejercicio de la ciencia y particularmente en lo que respecta a una adecuada transparencia en la transmisión de las investigaciones. Una experiencia en este sentido es el uso de insignias (badges) para señalar aquellas publicaciones que alcancen unos determinados criterios, como por ejemplo ofrecer los datos brutos en abierto (junto con el artículo) o compartir los códigos empleados para el análisis de datos (Kidwell et al., 2016). En el caso de la revista Psychological Science, por ejemplo, la introducción de estas insignias vino seguida de un incremento sustancial en la tasa de artículos que se acompañaban de datos o código en abierto, lo cual es un indicativo de que este tipo de estrategias de gamificación pueden ayudar a los científicos a superar sus reticencias iniciales a la transparencia.

Artículos de acceso abierto y sus ventajas sobre el impacto personal

Actualmente existe cierta diversidad en cuanto a las restricciones de acceso a la literatura científica publicada. Junto a revistas tradicionales (que cobran por el acceso a los artículos) conviven revistas de acceso abierto (algunas cargan el coste de la publicación al autor, otras no) y revistas híbridas (habitualmente, se trata de revistas de acceso por suscripción que puntualmente ofrecen a los autores la opción de liberar sus artículos). Dado que las publicaciones en abierto pueden ser visitadas por cualquier internauta interesado, sin necesidad de acceder a través de una entidad suscrita (como una universidad), cabe preguntarse si publicar en este formato podría de hecho incrementar la difusión de las investigaciones. En este sentido, se vienen realizando numerosas investigaciones que parecen confirmar esta ventaja de las publicaciones en abierto, al menos a corto-medio plazo (Gargouri et al., 2010): los artículos en abierto se leen más, se distribuyen en campos más heterogéneos y por ello acaban citándose más. Se trata, pues, de un nuevo incentivo que podría incrementar en el futuro la preferencia por publicar de forma abierta.

Incentivos al prerregistro: garantía previa de publicación

En secciones anteriores, se explicó cómo el sesgo de publicación (los resultados significativos que confirman la hipótesis planteada se publican con más probabilidad), junto con algunas prácticas cuestionables (p-hacking, HARKing, cherry picking y, en general, las derivadas de una excesiva flexibilidad del investigador) pueden introducir elementos distorsionadores en la práctica del científico (Gelman y Loken, 2014). Una estrategia que puede aliviar en cierta medida estos males es el prerregistro. Tradicionalmente, la revisión por pares de las revistas se efectúa sobre el artículo ya completo, con sus resultados y sus conclusiones. En un artículo prerregistrado, el primer envío consiste en un plan detallado del método y análisis de datos, que es evaluado por los revisores. Una vez pasado este filtro de revisión por pares, el estudio se lleva a cabo y el artículo se publica finalmente independientemente de cuál haya sido el resultado (e.g. Chambers, 2013). La garantía de publicación, sin importar si los datos resultan ser los esperados, es en sí mismo un aliciente para optar por esta modalidad de artículo que algunas revistas prestigiosas en nuestro campo ya están implantando (Psychological Science, Psychonomic Bulletin & Review).

Discusión

En este breve ensayo presentamos un panorama poco complaciente, pero creemos que realista, del estado actual de la credibilidad de la investigación en Psicología. El conjunto de problemas analizado no es exclusivo de la Psicología, sino que se extiende a otras ramas de las Ciencias Sociales, Biológicas y de la Salud (e.g., economía, medicina, nutrición, investigación en cáncer y neurociencia, entre otras sobre las que también se ha puesto el foco de forma particularmente intensa; Camerer et al., 2016; Errington et al., 2014; Prinz, Schlange y Asadullah, 2011). Sin embargo, el caso de la Psicología ha tenido una repercusión y un impacto mediático mayor. Esto, que podría considerarse injusto, se debe en parte a que la Psicología es una ciencia aún precaria, particularmente afectada por solapamientos con las pseudociencias, pero al mismo tiempo muy popular y de gran impacto en la política, la economía y la educación. El tamaño de la controversia es en gran medida el fruto del crecimiento de la importancia de la Psicología en la sociedad en las últimas décadas.

Desde nuestro punto de vista, que el foco se haya puesto en la Psicología puede servir de acicate para implantar las políticas de mejora que hemos comentado anteriormente. En ese sentido, podemos considerar positivo que la Psicología tome cartas en el asunto antes de que otras disciplinas ni tan siquiera hayan tomado conciencia del mismo. En la medida en que las soluciones sean efectivas en Psicología, pueden servir de ejemplo para las demás. Es, además, la Psicología la disciplina que puede aportar las herramientas aplicadas necesarias, pues es ella la ciencia que se preocupa por la relación entre los sistemas de incentivos, la cognición y la conducta.

En el momento actual, no podemos decir que el auto-escepticismo metodológico que defendemos aquí haya tenido un impacto generalizado sobre el comportamiento de los científicos. Es más, ese escepticismo se ha encontrado con una fuerte resistencia de una buena parte de los agentes que han contribuido al estado actual (Baumeister, 2016; Fiske, 2016). Gracias a la economía conductual, sabemos de la importancia del sesgo del statu quo y de la dificultad para corregirlo. También sabemos de la importancia de una aproximación constructiva, en la que el escepticismo no se convierta en un asunto de confrontación individual, sino en una herramienta psicoeducativa y de mejora de las estructuras en las que los agentes individuales tomas sus decisiones, siempre contingentes a unas circunstancias materiales determinadas.

Más allá de las medidas concretas, aún incipientes, que ya se están implantando en el sistema editorial, un signo para el optimismo es la corroboración de que la comunidad científica parece no navegar totalmente a ciegas en la niebla de ruido de datos que actualmente parece caracterizar a la Psicología. En algunos estudios recientes, se han establecido mercados de decisión para predecir si efectos concretos hallados en estudios científicos conseguirán replicarse y sobrevivir a un escrutinio posterior. Estos juegos, en los que los agentes invierten cantidades de dinero reales (de tal manera que están incentivados para acertar en sus decisiones de inversión individuales), han demostrado capacidad para predecir resultados políticos, económicos y sociales con bastante precisión (Surowiecki, 2004), bajo ciertas condiciones de operación y, sorprendentemente, han demostrado una capacidad superior al 70% de predecir la fiabilidad de un resultado científico (Dreber et al., 2015).

Dicho de otro modo, lejos de reducir los debates a pequeños comités de científicos interesados de forma intensa en problemas concretos (sometidos a incentivos, fuertemente polarizados y con un importante sesgo de statu quo, con los problemas que ellos conllevan), los datos sobre mercados de decisión sugieren que las predicciones son mejores cuando los agentes son independientes, están directamente incentivados para tomar decisiones objetivamente acertadas (y no por el sentido de la decisión grupal) y no son necesariamente expertos en la materia. Una buena parte de mejorar la ciencia psicológica pasa pues por democratizarla y someterla de forma abierta al escrutinio de la comunidad científica.

Referencias

Alberts, B. (2013). Impact Factor Distortions. *Science*, *340*(6134), 787-787.

American Society for Cell Biology. (2013). *The San Francisco Declaration on Research Assessment (DORA)*. Recuperado el 26 de Diciembre de 2016 de http://am.ascb.org/dora/

Anderson, M. S., Ronning, E. A., De Vries, R. y Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics, 13*(4), 437–461.

Arkes, H. R. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science, 22*, 356-360.

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. y Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, *7*, a1247.

Badenes-Ribera, L-. Frías-Navarro, D., y Pascual-Soler, M. (2015). Errors d'interpretació dels valors p en estudiants universitaris de psicologia. *Anuari de Psicologia, 16*(2), 15-31. doi: 10.7203/anuari.Psicologia.16.2.15

Baker, M. (2015, April). First results from psychology's largest reproducibility test. *Nature News*. Recuperado el 26 de Diciembre de 2016 de

http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*(8), 1069-1077.

Bakker, M., van Dijk, A. y Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bargh, J. A., Chen, M. y Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153-158.

Brembs, B., Button, K. y Munafò, M. (2013). Deep impact: unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*, a291.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. y Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews, 14,* 365-376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351,* 1433-1436.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610.

Chinn, C. A. y Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623-654.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*, 1304–1312.

Coyne, J. C., Thombs, B. D. y Hagedoorn, M. (2010). Ain't necessarily so: Review and critique of recent meta-analyses of behavioral medicine interventions in health psychology. *Health Psychology*, *29*(2), 107-116.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

DeKay, M. L. (2015). Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science, 24*, 405-411.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton y M. Borenstein (Eds.), *Publication bias and meta-analysis: Prevention, assessment and adjustments* (pp. 11-33). New York: John Wiley y Sons.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... y Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343-15347.

Enserink, M. (2012, November). Final Report: Stapel Affair Points to Bigger Problems in Social Psychology. *Science Insider*, Recuperado el 26 de Diciembre de 2016 de http://www.sciencemag.org/news/2012/11/final-report-stapel-affair-points-bigger-problems-social-psychology

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., y Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife, 3,* e04333.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A. y Baribault, B. (2016). *Understanding Bayes: How to become a Bayesian in eight easy steps: An annotated*

reading list. Recuperado el 26 de Diciembre de 2016 de https://alexanderetz.com/2016/02/07/understanding-bayes-how-to-become-a-bayesian-in-eight-easy-steps/

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS One*, *5*(4), e10271.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Fiske, S. T. (2016). A call to change science's culture of shaming. *APS Observer, 29*(9), 5-6.

Fons-Rosen, C. y Azoulay, P. (2015). *Does Science Advance One Funeral at a Time?* National Bureau of Economic Research, Working Paper No. 21788. Recuperado el 26 de Diciembre de 2016 de http://www.nber.org/papers/w21788.

Franco, A., Malhotra, N. y Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505.

Galton, F. (1907a). Vox populi (the wisdom of crowds). *Nature*, *75*, 450-451.

Galton, F. (1907b). One vote, one value. *Nature*, *75*, 414-414.

Gargouri, Y., Hajjem, C., Lariviére, V., Gingras, Y., Carr, L., Brody, T. y Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE*, *5*(10), e13636.

Gelman, A. y Loken, E. (2014). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Psychological Bulletin*, *140*(5), 1272–1280.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33(5), 587–606.

Higgins, J. P. y Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley-Blackwell.

Higginson, A. D. y Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, *14*(11), e2000995.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Hoekstra, R., Finch, S., Kiers, H. A. L. y Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*, 1033–1037.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645-654.

John, L. K., Loewenstein, G. y Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524 –532

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456. doi: 10.1371/journal.pbio.1002456

Kirschner, M. (2013). A Perverted View of "Impact." *Science*, *340*(6138), 1265-1265.

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... y Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "Many Labs" Replication Project. *Journal of Open Psychology Data, 2*(1), e4.

Lorenz, J., Rauhut, H., Schweitzer, F. y Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020-9025.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., Von Bastian, C. C., Carruth, N. P. y Miyake, A. (2016). No Evidence of the Ego-Depletion Effect across Task Characteristics and Individual Differences: A Pre-Registered Study. *PloS One*, *11*(2), e0147770.

Meehl, P. E. (1978) Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834

Myers, D. G. y Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627.

Nelson, N., Rosenthal, R. y Rosnow, R. L. (1986). Interpretation ofsignificance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299–1301.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

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

PLoS Blogs. (2015, February 25). *Positively Negative: A New PLOS ONE Collection focusing on Negative, Null and Inconclusive Results.* Recuperado el 26 de

Diciembre de 2016 de http://blogs.plos.org/everyone/2015/02/25/positively-negative-new-plos-one-collection-focusing-negative-null-inconclusive-results/

Prinz, F., Schlange, T., y Asadullah, K. (2011). Believe it or not: How much can we rely on publishd data on potential drug targets? *Nature Reviews Drug Discovery, 10,* 712.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S. y Weber, R. A. (2015). Assessing the Robustness of Power Posing No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, *26*(5), 653–656.

Rohrer, D., Pashler, H. y Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General, 144,* e73–e85.

Schervish, M. J. (1996). P-Values: What They Are and What They Are Not. *The American Statistician*, *50*(3), 203-206.

Sedlmeier, P. y Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., ... y Puhlmann, L. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives?. *Journal of Experimental Psychology: General*, *144*(6), e142-e158.

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

Smolčić, V. Š. (2013). Salami publication: definitions and examples. *Biochemia Medica*, *23*(3), 137–141.

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

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Anchor.

Szucs, D., y Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *bioRxiv*. doi: doi.org/10.1101/071530

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. y Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*, 6454-6459.

Van der Gucht, K., Griffith, J. W., Hellemans, R., Bockstaele, M., Pascal-Claes, F. y Raes, F. (2016). Acceptance and Commitment Therapy (ACT) for Adolescents: Outcomes of a Large-Sample, School-Based, Cluster-Randomized Controlled Trial. *Mindfulness*. Advance online publication. doi:10.1007/s12671-016-0612-y

Van Dijk, D., Manor, O. y Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*(11), R516–R517.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin y Review*, *14*(5), 779–804.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., y van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*, 1832. doi: 10.3389/fpsyg.2016.01832

Young, E. (2012, October). Nobel laureate challenges psychologists to clean up their act. *Nature News*. Recuperado el 26 de Diciembre de 2016 de http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535

Zatonski, T., y Witkowski, T. (2015). *Psychology Gone Wrong: The Dark Sides of Science and Therapy*. Boca Raton, FL, US: Brown Walker Publishers.

Zehetleitner, M. y Schönbrodt, F. (2016). *When does a significant p-value indicate a true effect? Understanding the Positive Predictive Value (PPV) of a p-value*. http://shinyapps.org/apps/PPV/