Revista Electrónica de Investigación en Filosofía y Antropología
NÚMERO 4 (Diciembre 2014)
Editor: Decanato de Filosofía. UNED
ISSN: 2340-4442

Alejandro Díaz García.
Departamento de Lógica, Historia y Filosofía de la Ciencia

Filosofía de la ciencia en acción: replicaciones, meta-análisis y *funnel plots* en psicología

Es un tema recurrente para la filosofía de la ciencia la distinción entre ciencia y pseudociencia. Entre aquello que puede considerarse conocimiento científico objetivo y aquello que no es más que especulación sin un correlato empírico. Frecuentemente el énfasis se ha puesto sobre la *falsabilidad* de los enunciados planteados por una determinada teoría, esto es, por la posibilidad de someterlos a refutación experimental. A su vez, la replicación de los resultados de una determinada observación empírica permite la acumulación de evidencia empírica. Y, especialmente, corroborar la veracidad, solidez y consistencia de dicho primer test experimental. La consistencia ha sido, asimismo, otro tema central para la evaluación de teorías científicas.

A lo largo de este artículo queremos incidir en el uso de métodos bibliográficos y estadísticos para el análisis de la práctica experimental en general, y psicológica en particular. Estos métodos exceden el uso de la lógica de enunciados para evaluar la consistencia de las predicciones y presupuestos de una determinada teoría científica. Pretenden mostrar, más allá de la consistencia de un determinado enfoque, fenómenos que aparecen en el contexto de la producción sucesiva de evidencia experimental y comunicación científica (el contexto de publicaciones). Hacemos referencia a la acumulación cuantitativa de conocimiento experimental y nuestra forma de integrarlo:

de qué forma lo hacemos disponible a través de la comunicación en publicaciones científicas de impacto -y cómo esta información queda estructurada-, y la forma en que hacemos meta-análisis, *i.e.*, análisis cuantitativos de variadas observaciones experimentales sobre un mismo tratamiento o hipótesis.

Algunos fenómenos asociados con dicha acumulación de evidencia empírica hacen referencia al *sesgo de publicaciones* y a la *limitada confiabilidad* de los resultados publicados. Intentaremos explicar en qué consisten estas consecuencias imprevistas de la práctica científica experimental, y cómo afectan a la calidad de la evidencia empírica disponible a través de los canales habituales de comunicación científica. Es decir, las publicaciones científicas y sus formas de visibilización en función de su impacto y relevancia.

1. Acumulando conocimiento científico: falsar hipótesis, replicar observaciones

El modelo intuitivo del método científico con el que cuenta un investigador experimental puede resumirse en comprobar hipótesis (probar un determinado tratamiento o manipulación experimental esperando que genere tal o cual efecto según las predicciones teóricas) y replicar resultados (o al menos el procedimiento que se espera que conduzca a dichos resultados). Esta lógica pretende materializar el principio popperiano de falsabilidad (cf. Popper, 1935). A esta lógica se le añaden los Test de Significatividad de la Hipótesis Nula (NHST, por sus siglas en inglés), esto es, los métodos de estadística inferencial que nos permiten estimar con qué probabilidad de error podemos dar una hipótesis por falsada (e.g., H(0): el psicofármaco X genera jaquecas), lo que se traduce en que afirmamos que un tratamiento es eficaz o seguro, que una determinada terapia es mejor que un placebo, así como otras consecuencias teóricas en la investigación básica. El resultado final es la producción de artículos experimentales que sirvan de resumen de dichas observaciones, y sus procedimientos experimentales (para así poder ser eventualmente replicadas), y recojan los parámetros estadísticos de dicho análisis.

Hasta aquí todo parece bien encajado en la estructura experimental. Sin embargo, ocasionalmente aparecen en la literatura experimental en psicología resultados contradictorios o totalmente incompatibles con las expectativas teóricas. Y dichos resultados parecen también poner en cuestión la estructura experimental desde la que se producen. Algunos de estos casos concretos tienen que ver con supuestas contrastaciones empíricas de telekinesis (e.g., Bem & Honorton, 1994) o fenómenos de *priming* conductual inconsciente[1] (e.g., Bargh, Chen & Burrows, 1996). Sin embargo, dado lo sorprendente de las observaciones, estos artículos ganaron una rápida difusión y generaron trabajo relacionado no necesariamente replicatorio ni refutatorio. ¿Por qué? A continuación mencionamos tres factores que creemos relevantes:

En primer lugar, algunos autores afirman que desde el punto de vista de las publicaciones, las replicaciones son poco atractivas (Pashler & Harris, 2012). En tanto que no constituyen investigación novedosa, la relevancia podría cuestionarse. De igual manera, es posible que replicar un fenómeno asumido como cierto dentro de la literatura haga que el proceso de revisión quede retrasado. Por ello, según argumentan Pasher y Harris, los autores no se ven atraídos a realizar replicaciones ya que suponen poco beneficio para su carrera investigadora.

Cabe distinguir en este punto entre replicaciones *directas* (o exactas), aquellas que reproducen exactamente el procedimiento experimental original para observar si se obtiene un resultado en el rango de lo esperable; de replicaciones *conceptuales*, aquellas que pretenden someter a prueba la misma hipótesis teórica general pero con una instanciación experimental distinta, lo que hace que se puedan introducir entre una y muchas variaciones en el diseño. Makkel (2012) ha contrastado que las replicaciones

1 El *priming* conductual (una especie de influencia inconsciente en el comportamiento) fue supuestamente demostrado, y ocasionalmente replicado, en un experimento en que los participantes, tras realizar unas sopas de letras en las que había frecuentes términos relacionados con la tercera edad (pensión, jubilación, Florida...), caminaban más despacio al salir del laboratorio que aquellos participantes del grupo control cuyas sopas de letras no tenían términos relacionados con la senectud. Se postulaba que una activación inconsciente de representaciones mentales de ancianos ralentizaba los movimientos de los participantes del grupo experimental. Varios experimentos recientes han sido incapaces de replicar los resultados (Doyen et al., 2012)

conceptuales son mucho más frecuentes en psicología que las directas (82 y 18% respectivamente). Esto puede deberse a la falta de atractivo antes mencionada que, evidentemente, afecta más a las replicaciones exactas en tanto que las replicaciones conceptuales puede probar la extensión de un fenómeno, presentar un nuevo diseño experimental, o probar variaciones estratégicas que muestren los límites de la predicción. Asimismo, las replicaciones conceptuales pueden ser vistas como un trabajo positivo (i.e., no explícitamente refutatorio), lo que puede favorecer la difusión de los resultados y el intercambio experimental. En este punto, algunos investigadores han señalado que privilegiar algunos temas (y evitar determinadas refutaciones) puede generar la aparición de burbujas científicas (Pedersen & Hendricks, 2013) por su capacidad para visibilizar a autores y generar temas que posibiliten beneficios competitivos dentro de la carrera académica o investigadora.

Segundo, la naturaleza misma de la investigación psicológica hace que algunas observaciones o procedimientos sean difíciles de reproducir. Es decir, que bajo supuestos como los de los programas neurolingüísticos, no podemos saber si una persona que está en una situación experimental generará siempre las mismas interacciones, pensamientos y respuestas; es decir, no sabremos qué replicar exactamente (cf. Bosman et. al., 2013). En este sentido, los experimentos que plantean constructos inobservables suelen tener un problema añadido para favorecer la aparición posterior de replicaciones.

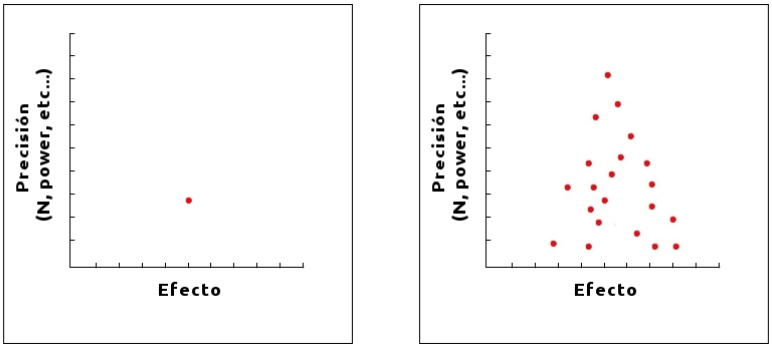
Por último, pero quizá más elemental aún, existe el problema derivado del uso del NHST que implica que los resultados no significativos (i.e., aquellos que no falsan la hipótesis nula) son de muy difícil interpretación teórica y suelen tacharse de inconcluyentes (Cohen, 1994). Esto, en un contexto de publicaciones que premia la relevancia y la novedad experimental, supone otra carga que hace que un resultado negativo (incluso aunque sea una replicación fallida) resulte poco interesante para la publicación, pues podría argumentarse que sólo es un resultado que no ha seguido correctamente los procedimientos, que no es una replicación suficientemente exacta, etc...

Cabe reseñar que el principio de caridad que se concede al experimentador cuando aporta un resultado positivo significativo, a menudo se torna en una exigencia de mejora cuando un resultado no lo es.

2. Agregación de resultados experimentales: *funnel-plots* y test de Egger.

A menudo la agregación de resultados de sucesivas observaciones experimentales de un mismo fenómeno/intervención, se realiza mediante estudios de meta-análisis que ponderan el poder estadístico de cada estudio y sus resultados para obtener un promedio que refleje el valor real más probable. Una forma de representar gráficamente la sucesiva acumulación de resultados son los llamados *funnel plots*, que sitúan cada resultado experimental en función de dos dimensiones: el efecto y el poder del estudio (indicado por su tamaño muestral (N), su poder (P) u otra medida de calidad del estudio).

En la figura 1 podemos ver una progresión ideal de acumulación de conocimiento al respecto de una intervención concreta. Una primera observación con un poder estadístico medio se va viendo corroborada posteriormente por observaciones más o menos potentes que generan resultados que tienden a distribuirse de manera simétrica y piramidal: a la base hay abundantes estudios de limitado poder estadístico con una gran variabilidad de resultados y en la parte alta vemos muy pocos resultados, en tanto que realizar estudios con más poder estadístico es más trabajoso, y que su efecto suele ser -aproximadamente- el valor medio del resto de observaciones.



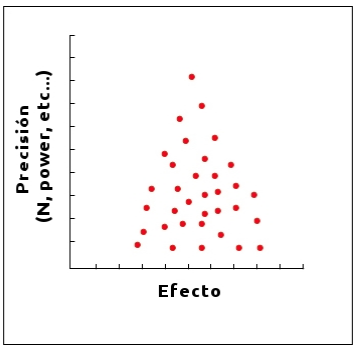


Figura 1.- Representación esquemática ideal de la acumulación de resultados de observaciones experimentales.

Esta forma gráfica de representar los resultados nos permite obtener información visual de cómo se lleva a cabo la sucesiva experimentación sobre un determinado fenómeno. Ahora imaginemos que, tal y como hemos mencionado anteriormente, los resultados no significativos (aquellos con un efecto menor en este caso) son de difícil interpretación y, por tanto, su publicación es más trabajosa. Si existiera un sesgo de publicación hacia los resultados no significativos (además de hacia aquellos producidos con poco poder estadístico) la literatura resultante generaría un patrón más parecido al de la figura 2.

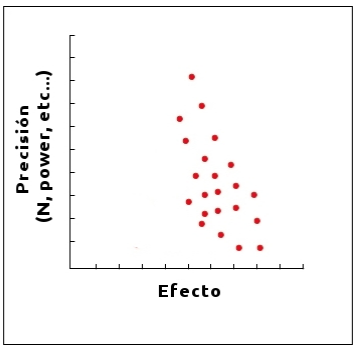


Figura 2.- Representación esquemática de la literatura disponible sobre un determinado fenómeno/intervención cuando existe sesgo de publicación.

Como podemos observar, los resultados que no alcanzan un cierto tamaño del efecto (a menudo determinado por el nivel de significatividad) no llegan a ser publicados. La lógica que subyace es que los autores que los obtienen bien ven como son rechazados por inconcluyentes, o bien deciden dejarlos apartados en un archivo del laboratorio (i.e., el llamado efecto *file-drawer*; Rosenthal, 1979) ya que saben que publicarlos puede suponer mucho más esfuerzo que intentan realizar otra serie de experimentos. La consecuencia resulta intuitiva: los estudios de meta-análisis quedan

sesgados por este tipo de fenómenos. No es difícil imaginar que la media observada tras un meta-análisis para la intervención que recoge la figura 2 estará sesgada con respecto a la de la figura 1, dando sensación de que los efectos de esa terapia son mayores de lo que realmente son.

Sin embargo, además de la visión intuitiva que nos dan los *funnel plots*, hay indicadores analíticos que nos permiten medir el sesgo de publicaciones para un determinado problema, tratamiento o intervención. En general son medidas de simetría sobre la distribución de resultados a través de diferentes experimentos, y miden si los efectos se asocian con varianzas altas (es decir, si se ha explotado el uso de casos que pudieran ser artefactos y se los ha incluido en el estudio por su valor extremo que puede incrementar la significatividad). Dos de los más usados son el test de Begg y el test de Egger (cf. Sterne, Egger & Davey-Smith, 2001).

3. Un poco más sobre el sesgo de publicación

Desafortunadamente, el sesgo de publicación es algo en lo que el investigador aplicado no siempre repara a la hora de evaluar los resultados presentes en un determinado estudio, en un meta-análisis o en una revisión de la literatura. De hecho los métodos para estudiarlo no suelen formar parte de su formación académica. Y sólo recientemente estos estimadores comienzan a presentarse de manera sistemática cuando se realizan artículos y estudios de meta-análisis (e.g., Virués-Ortega, 2010).

Sin embargo, su incidencia en el contexto académico se ha puesto especialmente de manifiesto en los últimos años. Existen inquietudes entre algunos autores de que la creciente competición por obtener un volumen suficiente de publicaciones, citas y financiación pueda distorsionar la labor experimental.

Daniele Fanelli (2012) ha analizado más de 4600 artículos publicados entre 1990 y 2007 en las principales publicaciones de todas las áreas científicas. Lo que observó, curiosamente, fue que los resultados negativos (i.e., aquellos no estadísticamente

significativos, los que no demostraban la hipótesis experimental) estaban desapareciendo progresivamente de la literatura. De hecho, el esquema actual de publicaciones hace que los artículos normalmente se formulen en términos de "evidencia a favor" de una determinada hipótesis o tratamiento. Evidentemente, esta no era una predicción esperable de la lógica de la investigación. Es verdad que algunas técnicas modernas de recogida de datos pueden favorecer métodos estadísticos más exhaustivos y técnicas de recogida de datos que aumenten el volumen de observaciones (lo que genera que, de haber un efecto real, sea más fácil obtener una muestra experimental en que el resultado sea significativo). Pero parece que este fenómenos tiene más que ver con la dificultad de publicar resultados negativos. Esta tendencia en los últimos años está presente en casi todas las áreas y regiones geográficas.

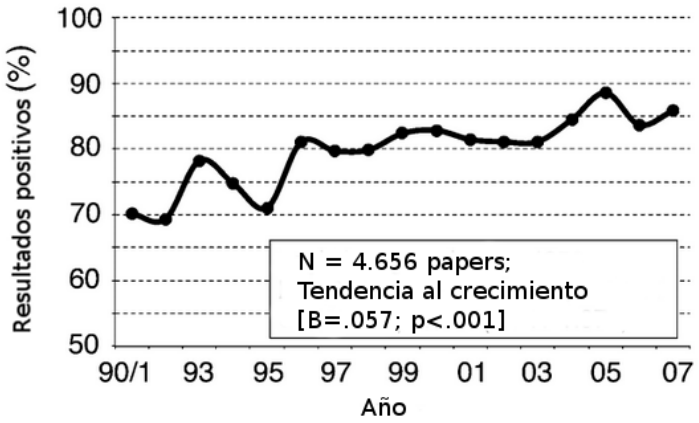


Figura 3.- Ratio de artículos en la literatura que muestran evidencia a favor de la hipótesis experimental en función del año.
Adaptado a partir de Fanelli (2012)

La tendencia que podemos observar en la figura 3 es generalmente más acusada en ciencias sociales con respecto a aquellas llamadas ciencias duras. Pero algunas áreas concretas de estas últimas son también ejemplos notorios de desaparición de resultados negativos (i.e., biología molecular, farmacología y ensayos clínicos en medicina), mientras que otras áreas como las ciencias ambientales, las ciencias aeronáuticas o la física parecen tener un ratio estable. En la figura 4 podemos observar la tendencia en psicología con una muestra de 277 artículos. La tendencia es significativamente creciente (B =0.154, p=.004).

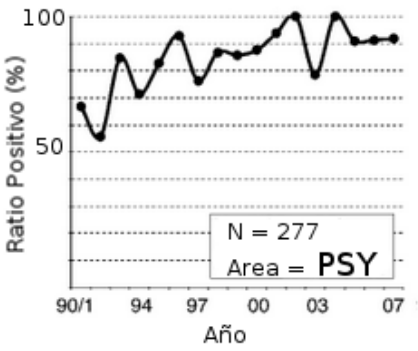


Figura 4.- Proporción de resultados positivos en psicología en función del año. Adaptado a partir de Fanelli (2012).

Cabe mencionar que este patrón se repite en todas las áreas geográficas, hecho indicativo de un mundo académico interdependiente. La tendencia es más sostenida en Estados Unidos, mientras que Asia ha acostumbrado, tradicionalmente, a tener mayor proporción de resultados positivos. Asimismo, el propio Fanelli (2011) ha observado que en aquellos estados de los EE. UU. con más gasto en investigación también hay una mayor proporción de resultados positivos.

4. Conclusiones

Como hemos visto la práctica científica no siempre se organiza de forma espontánea encarnando los supuestos de la supuesta lógica del descubrimiento y la justificación científica. En la medida en que es una actividad humana orientada a la comunicación y el intercambio, está sujeta a que otros fenómenos aparezcan e interactúen con el esperado desarrollo del conocimiento y la práctica experimental. Hemos querido mostrar en estas páginas que existen técnicas analíticas para abordar el estudio de la producción científica y el testeo de hipótesis, que van más allá de la tradicional evaluación de consistencia de un determinado enunciado teórico con otros enunciados teóricos y/o empíricos.

Sin embargo, a pesar de las buenas noticias que dichas técnicas suponen, este artículo también pretender reivindicar una lectura crítica de la literatura científica. Es decir, reclamar que ésta debe entenderse como producto del contexto social (la carrera

académica) en el cual se produce. Asimismo, las distintas áreas de conocimiento aportan particularidades que hacen que la interpretación de cada ejercicio experimental pueda ser más o menos abierta, y constituir, en mayor o menor grado, un verdadero ejercicio de refutación o falsación.

Ya en el ámbito concreto de la psicología, hemos de mencionar una de las consecuencias fundamentales de la estructuración de la práctica científica. Y es, tal y como algunos autores han caracterizado, la existencia de un enorme cementerio de teorías zombie (Ferguson & Heene, 2012). Es decir, de principios, teorías e hipótesis que nunca terminan de ser falsadas o no permiten nunca una refutación completa, debido a la naturaleza misma del fenómeno o de nuestros constructos explicativos. Esta dificultad para falsar teorías puede incrementar la aparición de burbujas científicas o temas candentes que pueden generar beneficios en la carrera investigadora individual sin generar un verdadero consenso científico o una revolución paradigmática (cf. Kuhn, 1962).

Es por ello que esperamos que este artículo sirva para defender una lectura crítica de toda evidencia, para poder considerarla fruto no sólo de una tradición teórico-experimental, sino fruto de dinámicas sociales en un sistema que, paradójicamente, es simultáneamente colaborativo y competitivo.

Referencias

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

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin, 115*(1), 4-18.

Bosman, A. M., Cox, R. C., Hasselman, F., & Wijnants, M. L. (2013). From the Role of Context to the Measurement Problem: The Dutch Connection Pays Tribute to Guy Van Orden. *Ecological Psychology, 25*, 240-247.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PloS One, 7*(1), e29081.

Fanelli, D. (2010) Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLoS One 5*(4), e10271.

Fanelli, D. (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561.

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

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research How Often Do They Really Occur?. *Perspectives on Psychological Science, 7*, 537-542.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531-536.

Pedersen, D. B., & Hendricks, V. F. (2014). Science Bubbles. *Philosophy & Technology, 27*, 503-518.

Popper, K. R. (1935/2008). *La lógica de la investigación científica.* Ed. Tecnos.

Sterne J.A.C., Egger M. & Davey-Smith G. (2001). Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal, 323*, 101-105.

Virués-Ortega, J. (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical psychology review, 30*, 387-399.