El manifiesto de Leiden sobre indicadores de investigación

Diana Hicks[a], Paul Wouters[b], Ludo Waltman[b], Sarah de Rijcke[b] and Ismael Rafols[c,d,e]

[a] School of Public Policy, Georgia Institute of Technology, Atlanta, USA
[b] Centre for Science and Technology Studies (CWTS), University of Leiden, The Netherlands
[c] *Ingenio* (CSIC-UPV), Universitat Politècnica de València, València, Spain
[d] Science Policy Research Unit (SPRU), University of Sussex, Brighton, UK
[e] Observatoire des Science et des Técniques (OST-HCERES), Paris, France

(Traducción al castellano de Hicks et al. (2015) The Leiden Manifesto for research metrics. *Nature*, 520, 429-431. www.ingenio.upv.es/manifiesto)

Los datos sobre las actividades científicas están siendo cada vez más utilizados para gobernar la ciencia. Evaluaciones sobre investigación que fueron en su día diseñadas individualmente para su contexto específico y realizadas por pares, son ahora rutinarias y están basadas en métricas.[1] El problema es que la evaluación pasó de estar basada en valoraciones de expertos a depender de estas métricas. Los indicadores han proliferado: normalmente bien intencionados, no siempre bien informados, y a menudo mal aplicados. Cuando organizaciones sin conocimiento sobre buenas prácticas e interpretación apropiada de indicadores llevan a cabo las evaluaciones, corremos el riesgo de dañar el sistema científico con los mismos instrumentos diseñados para mejorarlas.

Antes del año 2000, los expertos utilizaban el Science Citation Index del Institute for Scientific Information (ISI), en su versión de CD-ROM para realizar análisis especializados. En el 2002, Thomson Reuters lanzó una plataforma web integrada que hizo accesible a un público amplio la base de datos Web of Science. Luego aparecieron otros índices de citas que se erigieron en competencia de Web of Science: Scopus de Elsevier (2004) y Google Académico (versión beta creada en el 2004). Instrumentos basados en la web fueron luego introducidos, tales como InCites (que usa Web of Science) y SciVal (que usa Scopus) y también software para analizar perfiles individuales de citas basados en Google Académico (Publish or Perish, que apareció el 2007).

En el 2005, Jorge Hirsch, un físico de la Universidad de California en San Diego, propuso el índice-*h,* que popularizó el recuento de citas de investigadores individuales. El interés en el factor de impacto de las revistas académicas creció incesantemente desde 1995. Recientemente, han aparecido medidas de uso social y de comentarios *online*: F1000Prime fue establecido en 2002, Mendeley en 2008 y Altmetric.com en 2011.

En tanto que investigadores de cientometría, científicos sociales y gestores de investigación, hemos observado con creciente preocupación un uso incorrecto generalizado de los indicadores en la evaluación del desempeño científico. Los siguientes son algunos de los numerosísimos ejemplos posibles. En todo el mundo, las universidades se han obsesionado con su posición en los rankings globales (tales como el ranking de Shanghai y la lista del *Times Higher Education*), cuando estas listas están basadas en lo que a nuestro juicio son datos inexactos e indicadores arbitrarios.

[1] Wouters, P. in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).

Algunas organizaciones piden el índice-*h* a los candidatos que se presentan a ofertas de empleo. Varias universidades basan la promoción en valores umbral del índice-*h* y en el número de artículos en revistas de "alto impacto". Los CVs se han convertido en oportunidades de alardear de estas "puntuaciones", en particular en biomedicina. En todas partes, los supervisores piden prematuramente a sus estudiantes de doctorado que publiquen en revistas de alto impacto y consigan financiación externa.

En Escandinavia y China, algunas universidades distribuyen fondos de investigación o bonificaciones sobre la base de un número: por ejemplo, calculando puntaciones individuales de impacto para repartir "recursos de desempeño", o dando a los investigadores una prima por publicaciones en una revistan con un factor de impacto superior a 15.[2]

Por estas razones, presentamos el *Manifiesto de Leiden*, que recibe este nombre de la conferencia donde cristalizó (http://sti2014.cwts.nl). Sus diez principios no son ninguna novedad para expertos en cientometría, pero ninguno de nosotros sería capaz de recitarlos en su totalidad puesto que hasta este momento no habían sido codificados. Celebridades en cientometría, como Eugene Garfield (fundador de ISI), ya han presentado en ocasiones algunos de estos principios,[3] pero no pueden estar presentes cuando los evaluadores informan a gestores universitarios que no son expertos en la metodología pertinente. Los científicos que buscan literatura para disputar o impugnar evaluaciones sólo encuentran las informaciones necesarias en lo que son, para ellos, revistas opacas y de difícil acceso.

Ofrecemos esta síntesis de buenas prácticas en evaluación basada en indicadores métricos para que los investigadores puedan pedir cuentas a los evaluadores, y para que los evaluadores puedan pedir cuentas a los indicadores.

DIEZ PRINCIPIOS

1. La evaluación cuantitativa tiene que apoyar la valoración cualitativa por expertos.

Los indicadores pueden corregir la tendencia a perspectivas sesgadas que se dan en revisión por pares y facilitar la deliberación. En este sentido, los indicadores pueden fortalecer la evaluación por pares puesto que tomar decisiones sobre colegas es difícil sin varias fuentes de información. Sin embargo, los evaluadores no deben ceder a la tentación de supeditar las decisiones a los números. Los indicadores no pueden sustituir a los razonamientos informados. Los decisores tienen plena responsabilidad sobre sus evaluaciones.

2. El desempeño debe ser medido de acuerdo con las misiones de investigación de la institución, grupo o investigador.

Los objetivos de un programa de investigación tiene que ser especificados al principio, y los indicadores usados para medir el desempeño tienen que estar claramente relacionados con estos objetivos. La elección y usos de los indicadores tiene que tener en cuenta los contextos socio-económicos y culturales. Los científicos tienen diversas misiones de investigación. La investigación

[2] Shao, J. & Shen, H. *Learned Publishing* **24,** 95–97 (2011).
[3] Seglen, P. O. *Br. Med. J.* **314,** 498–502 (1997). Garfield, E. *J. Am. Med. Assoc.* **295,** 90–93 (2006).

para avanzar las fronteras del conocimiento académico es diferente de la investigación focalizada en proveer soluciones a problemas sociales. La evaluación puede estar basada en méritos relevantes para la industria, el desarrollo de políticas, o para los ciudadanos en general, en vez de méritos basados en nociones académicas de excelencia. No hay un modelo de evaluación que se pueda aplicar en todos los contextos.

3. La excelencia en investigación de relevancia local debe ser protegida.

En muchas partes del mundo, excelencia en investigación se asocia únicamente con publicaciones en inglés. La ley española, por ejemplo, explicita el deseo y la conveniencia que los académicos españoles publiquen en revistas de alto impacto. El factor de impacto se calcula para revistas indexadas por Web of Science, que es una base de datos basada en los Estados Unidos y que contiene una gran mayoría de revistas en inglés. Estos sesgos son especialmente problemáticos en las ciencias sociales y las humanidades, áreas en las que la investigación está más orientada a temas regionales y nacionales. Muchos otros campos científicos tienen una dimensión nacional o regional -- por ejemplo, epidemiología del VIH en el África subshariana.

Este pluralismo y la relevancia social tienden a ser suprimidos cuando se crean artículos de interés a los guardianes del alto impacto: las revistas en inglés. Los sociólogos españoles muy citados en Web of Science han trabajado en modelos abstractos o estudiado datos de los Estados Unidos. En ese proceso se pierde la especificidad de los sociólogos con alto impacto en las revistas en castellano: temas como la ley laboral local, atención médica para ancianos o empleo de inmigrantes.[4] Indicadores basados en literatura de alta calidad no inglesa servirían para identificar y recompensar la excelencia en investigación localmente relevante.

4. Los procesos de recopilación y análisis de datos deben ser abiertos, transparentes y simples.

La construcción de las bases de datos necesarias para evaluar debe seguir procesos establecidos antes de que la investigación sea completada. Ésta ha sido la práctica común entre los grupos académicos y comerciales que han desarrollado metodologías de evaluación durante varias décadas. Estos grupos publicaron los protocolos de referencia en la literatura revisada por pares. Esta transparencia permite el escrutinio y control de los métodos. Por ejemplo, en 2010, un debate público sobre las propiedades técnicas de un importante indicador utilizado por uno de nuestros grupos (el Centro de Estudios de Ciencia y Tecnología (CWTS) de la Universidad de Leiden, en los Países Bajos), se saldó con una revisión en el cálculo de este indicador.[5] Las nuevas empresas comerciales en el campo deben responder a los mismos estándards. Nadie tiene porque aceptar evaluaciones automáticas salidas de caja negras o procesos impenetrables. La simplicidad es una virtud en un indicador porque favorece la transparencia. Pero indicadores simplísticos pueden distorsionar la evaluación (veáse el principio 7). Los evaluadores debe esforzarse en encontrar un equilibrio: indicadores simples que sea respetuosos con la complejidad de los procesos de investigación descritos.

5. Los datos y análisis deben estar abiertos a verificación por los evaluados

[4] López Piñeiro, C. & Hicks, D. *Res. Eval.* 24, 78–89 (2015).
[5] van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. *J. Informetrics* 4, 431–435 (2010).

Con el fin de asegurar la calidad de los datos, los investigadores incluídos en estudios bibliométricos tienen que poder comprobar que sus contribuciones han sido correctamente identificadas. Los responsables y gestores de los procesos de evaluación deben garantizar la exactitud de los datos usados mediante métodos de auto-verificación o auditoría por terceras partes. La universidades podrían implementar este principio en sus sistemas de información. Este debería ser un principio rector en la selección de proveedores de estos sistemas. La compilación y proceso de datos de alta calidad, precisos y rigurosos, lleva tiempo y cuesto dinero. Los responsables deben asignar presupuestos a la altura de estas necesidades de calidad.

6. Las diferencias en las prácticas de publicación y citación entre campos científicos deben tenerse en cuenta.

La mejor práctica en evaluación es proponer una batería de indicadores y dejar que los distintos campos científicos escojan los indicadores que mejor les representan. Hace unos años, un grupo de historiadores recibió una puntuación relativamente baja en una evaluación nacional de pares porque escribían libros en vez de artículos en revistas indexadas por Web of Science. Estos historiadores tuvieron la mala suerte de formar parte del departamento de psicología. La evaluación de historiadoes y científicos sociales requiere la inclusión de libros y literatura en la lengua local; la evaluación de investigadores en informática necesita considerar las contribuciones a conferencias.

La frecuencia de citación varía según los campos: las revistas más citadas en ránkings de matemáticas tienen un factor de impacto alrededor de 3; las revistas más citadas en ránkings de biología celular tienen factors de impactor alrededor de 30.

Por lo tanto, se necesitan indicadores normalizados por campo, y el método más robusto de normalización esta basado en percentiles: cada publicación es ponderada según el percentil al que pertenece en la distribución de citaciones de su campo (por ejemplo, el percentil 1%, 10%, 20% más alto). Una única publicación altamente citada mejora un poco la posición de una universidad en un ranking basado en percentiles, pero puede propulsar la universidad de un lugar medio a la primeras posiciones en un ranking basado en promedios de citas.[6]

7. La evaluación individual de investigadores debe basarse en la valoración cualitativa de su portafolio de investigación.

El índice-*h* aumenta con la edad del investigador, aunque éste ya no publique. El índice-*h* varía por campos: los científicos en las ciencias de la vida pueden llegar a 200; los físicos a 100 y los científicos sociales a 20 o 30.[7] Es un índice que depende de la base de datos: hay informáticos que tienen un índice-*h* de 10 en Web of Science, pero de 20 o 30 en Google Scholar.[8] Leer y valorar el trabajo de un investigador es mucho más apropiado que confiar en un único número. Incluso cuando se comparan un gran número de científicos, es mejor adoptar un enfoque que considere información diversa sobre cada individuo, incluyendo sus conocimiento, experiencia, actividades e influencia.

8. Debe evitarse la concreción improcedente y la falsa precisión.

[6] Waltman, L. *et al. J. Am. Soc. Inf. Sci. Technol.* 63, 2419–2432 (2012).
[7] Hirsch, J. E. *Proc. Natl Acad. Sci. USA* 102, 16569–16572 (2005).
[8] Bar-Ilan, J. *Scientometrics* 74, 257–271 (2007).

Los indicadores de ciencia y tecnología tienden a la ambigüedad conceptual y a la incertidumbre, y se fundamentan en hipótesis que no están universalmente aceptadas. Por esta razón, las buenas prácticas usan múltiple indicadores con el fin de construir un retrato robusto y plural. En la medida que sea posible cuantificarla, información sobre incertidumbre y error debería acompañar la valores de los indicadores publicados, por ejemplo usando barras de error. Si esto no fuera posible, los productores de indicadores deberían al menos evitar ofrecer un falso nivel de precisión. Por ejemplo, el factor de impacto de revistas se publica con tres decimales para evitar empates. Sin embargo, dada la ambigüedad conceptual y la variabilidad aleatoria de las citas, no tiene sentido distinguir entre revistas por pequeñas diferencias en el factor de impacto. Se debe evitar la falsa precisión: sólo un decimal está justicaficado.

9. Deben reconocerse los efectos sistémicos de la evaluación y los indicadores.

Los indicadores cambian el sistema científico a través de los incentivos que establecen. Estos efectos deberían ser anticipados. Esto significa que una batería de indicadores es siempre preferible puesto que un solo indicador es susceptible de generar comportamientos estratégicos y substitución de objetivos (según la cual la medida se convierte en un fin en sí misma). Por ejemplo, en los 1990s, Australia financió investigación en universidades de acuerdo con una fórmula basada sobretodo en el número de publicaciones de un instituto. Las universidades podían calcular el "valor" de una publicación en una revista arbitrada; en el año 2000, el valor se estimó en Aus$800 (US$480) destinados a recursos de investigación. Como era de esperar, el número de artículos publicados por autores australianos subió, pero en revistas menos citadas, lo que sugiere que la calidad de los artículos disminuyó.[9]

10. Los indicadores deben ser examinados y actualizados periódicamente.

Las funciones de la investigación y los objetivos de la evaluación cambian o se desplazan, y el sistema de investigación co-evoluciona con ellos. Medidas que fueron útiles en su día pasan a ser inadecuadas y nuevos indicadores aparecen. Por lo tanto, los sistemas de indicadores tienen que ser revisados y tal vez modificados. Al darse cuenta de los efectos de su fórmula simplista de evaluación, en 2010 Australia adoptó la iniciativa Excellence in Research for Australia, que es más compleja y pone énfasis en la calidad.

Pasos siguientes

Siendo fiel a estos diez principios, la evaluación de la investigación puede jugar un papel importante en el desarrollo de la ciencia y sus interacciones con la sociedad. Los indicadores de investigación pueden proporcionar información crucial que sería difícil de aglutinar o entender a partir de experiencias individuales. Pero no se debe permitir que la información cuantitativa se convierta en un objetivo en sí misma. Las mejores decisiones se toman combinando estadísticas robustas sensibles a los objetivos y la naturaleza de la investigación evaluada. Tanto la evidencia cuantitativa como la cualitativa son necesarias -- cada cual es objetiva a su manera. Decisiones sobre la ciencia tienen que ser tomadas en base a procesos de alta calidad informados por datos de la mayor calidad.

[9] Butler, L. *Res. Policy* 32, 143–155 (2003).