Ciencia bajo observación.
Beneficios, límites y paradojas de la evaluación de la actividad científica

Daniel Innerarity

Catedrático de Filosofía Política y Social, Profesor de Investigación "Ikerbasque" en la UPV/EHU y Director del Instituto de Gobernanza Democrática

La ciencia es algo demasiado importante como para dejarla únicamente en manos de los científicos. En las sociedades democráticas el principio de autonomía de la ciencia es complementado con el de responsabilidad. La actividad científica se desarrolla hoy en una sociedad que es cada vez más exigente respecto a las responsabilidades de los científicos; esta vigilancia pública se refiere tanto a las grandes cuestiones que plantean profundos dilemas éticos como a la supervisión acerca de qué se hace con los recursos públicos. La gobernanza del conocimiento se encuentra en medio de está tensión: ¿cómo compaginar su responsabilidad con el hecho de que la ciencia requiere un espacio de autogobierno y libertad que no esté sometido a los imperativos inmediatos de la relevancia social o la rentabilidad económica?

Esta es la razón que está detrás multitud de discusiones acerca de qué es lo que podemos esperar de la ciencia y en qué medida debe estar sometida a un control social, compatible con su principio de autogobierno, sin retroceder en la conquista de un espacio de libertad e innovación fuera del cual no es pensable el ejercicio de la actividad científica. Toda una serie de instituciones y procedimientos, entre ellos el de su evaluación, tienen precisamente como objetivo asegurar el cumplimiento de las expectativas que la sociedad deposita en la ciencia, de manera cada vez más creciente. Se trata de un conjunto de exigencias de legitimación en decisiones que nos

afectan a todos, en la asignación de recursos o en la regulación del hecho de que haya cada vez más autores y más fuerzas que intervienen en asuntos que ya no pueden considerarse como una competencia exclusiva de los expertos.

Esta exigencia de rendición de cuentas no es algo que se dirija únicamente a la ciencia, sino que se ha constituido en un principio normativo central de nuestras sociedades. El diagnóstico social que está detrás de denominaciones como *"audit society"* (Power 1977), *"age of inspection"* (Day/Klein 1990), *"evaluative state"* (Neave 1988), *"accounting society"* (Weingart 2005) se refiere a una tendencia general a configurar estructuras de "control del control" en los diversos subsistemas de la sociedad contemporánea.

Me propongo llamar aquí la atención sobre los límites y las paradojas de la evaluación, como procedimiento que debe ser utilizado reflexivamente, es decir, valorando también sus costes, sus posibles errores e incluso sus efectos indeseados. De hecho, generalmente cuando se hace una evaluación la cuestión de su coste no es suficientemente tematizada. Esto tiene como consecuencia el hecho de que su utilidad es sistemáticamente supravalorada. Quisiera subrayar especialmente los costes epistemológicos de un sistema que también puede producir una "dinámica de la desconfianza en la era de la calidad" (Sitkin /Stickel 1996), un efecto que las ciencias sociales llevan tiempo investigando y que algunos han llamado "transintencionalidad" (Greshoff / Knerr / Schimank 2003), es decir, el conjunto de efectos no deseados que acompañan a cualquier tarea de gobierno y organización: la introducción de controles con la intención de fortalecer el control y la transparencia dentro de las organizaciones puede tener el efecto contrario.

Las expectativas de racionalización que impulsan el despliegue de las evaluaciones no dejan de ser problemáticas. Por una parte, pueden ser expectativas desmesuradas en relación con los instrumentos de la evaluación, pero también pueden hacer que se desdibuje el objetivo mismo de la evaluación. ¿Qué es lo que estamos pretendiendo: significación epistémica, relevancia social o adjudicación

eficiente de recursos? Es una ilusión pensar que estas valoraciones van siempre unidas o que no plantean nunca problemas de compatibilidad.

Aunque en buena parte de las críticas a la evaluación tienen un fondo de conservadurismo o defienden el *status quo,* la reflexión acerca de su sentido, límites y paradojas puede contribuir no tanto a abolirla como a mejorarla. Esta reflexión de segundo grado —evaluar la evaluación— es necesaria, de entrada, porque no podemos actuar como si los expertos no se equivocaran nunca; nuestros procedimientos de evaluación deberían estar siempre alertas ante tal posibilidad. ¿Cómo protegernos de quiebras de la confianza en el sistema científico similares a la que ha sufrido la economía con el descrédito de un instrumento de mediación tan importante como las agencias de *rating?*

Propongo compendiar estas limitaciones en cuatro tipos: las limitaciones de la exactitud, lo que podríamos llamar los efectos secundarios de la evaluación, la dificultad de evaluar lo realmente nuevo en la ciencia y finalmente las limitaciones que provienen de su falsa universalidad.

1. Cantidad y calidad

La evaluación tiene la aspiración de medir la calidad y ambos términos —medición y calidad— no siempre son compatibles. No me refiero a cosas que no estén siendo suficientemente medidas sino a lo que no se deja medir por su propia naturaleza, a los límites intrínsecos de la exactitud, a la dificultad de traducir lo cuantitativo en un juicio de calidad y permitir así el *"governing by numbers"* (Porter 1995; Miller 2001).

La seducción de lo cuantitativo ha puesto en marcha una bibliometría que gestiona cantidades, números, citas e impactos de manera que evaluadores y evaluados orientemos nuestros respectivos trabajos más por la cantidad que por la

calidad. El problema es que la bibliometría no permite distinguir al verdadero investigador del hábil profesional, los trabajos innovadores de las publicaciones rutinarias en serie, la hiperproductividad mediocre de la investigación eficiente (Hornbostel 1997). Sus limitaciones como método de gobernanza de la ciencia son patentes en cuanto analizamos el alcance de instrumentos de medida como la cita o el análisis del impacto.

De entrada, la cita tiene una diversa significación en cada una de las disciplinas científicas. Cada disciplina tiene una peculiar cultura de la cita; las citas pueden ser positivas, negativas o simplemente rituales. Las excelencia de esta investigación y su transferencia a la sociedad no es bien ponderada cuando se valora únicamente por las citas en revistas internacionales especializadas. Si valoramos la interdisciplinaridad, las citas provenientes de otros campos deberían tener un mayor valor que las de nuestra disciplina. El impacto se mide por las veces que es citado, con independencia de que sea para criticarlo o alabarlo. Existen ciertos "clubs de citas" auto-referenciales y bastantes evidencias de que, por ejemplo, los científicos norteamericanos solo se citan entre sí.

Algo similar puede decirse de los *rankings* y su aceptación irreflexiva. Los números con los que se clasifica a los investigadores, a las universidades y a los centros de investigación no son medidores neutros y objetivos; por el contrario, contienen juicios previos, deben ser interpretados y no pueden ser utilizados ilimitadamente. La cuestión acerca de qué debe entenderse por excelencia o calidad en relación con el saber es algo que no puede confiarse únicamente a clasificaciones e indicadores indiscutibles sino que exigen un permanente debate público sobre la política del conocimiento. Otro indicador puede ser la consecución de fondos para la investigación, que no dice mucho acerca del sentido o la productividad de la investigación financiada. Si está muy extendido este y otros procedimientos de medida es simplemente porque son cantidades fáciles de medir.

Las dificultades de traducir lo cuantitativo en cualitativo no impiden el empleo de métodos cuantitativos; es más bien una llamada de atención contra cierta beatería de los números, la ilusión de la exactitud o el culto de la cantidad. Las mediciones cualitativas, cuando están bien hechas, facilitan el juicio y la toma de decisiones pero no las hacen innecesarias.

2. Los efectos secundarios de la evaluación

Cuando los seres humanos son observados reaccionan modificando su conducta. El instrumento de la evaluación modifica de modo sistemático aunque no intencional el comportamiento de las personas afectadas. Las ciencias del comportamiento han estudiado desde hace años esta tendencia de las personas a concentrarse exclusivamente en los criterios medidos y desatenderse de todo lo demás. La atención a ciertos indicadores motiva a los investigadores a desarrollar una competitividad que se traduce, por ejemplo, en la cantidad de las publicaciones; algunos las aumentan de tal modo que trocean sus trabajos hasta las más pequeñas unidades publicables (Tucci 2006, 28); otros proponen proyectos de investigación menos innovadores pero más seguros; los hay que tienden a disminuir la exigencia de sus doctorados.

El principal efecto secundario de la evaluación es que promueve indirectamente una fuerza de adaptación en virtud de la cual la motivación principal ya no es la búsqueda de nuevo saber sino con qué actividades, temas y productos se consigue más fácilmente una evaluación positiva. La publicación se convierte en un fin en sí mismo. Cuando reina este juego de inercias podemos estar seguros de que se está produciendo ese efecto de control en virtud del cual las motivaciones intrínsecas del trabajo disminuyen en la medida en que aumentan los incentivos externos (Bénabou / Tirole 2003; Fehr / Gächter 2002).

Sabemos ya mucho sobre los límites de la medida y el juego de una *"creative compliance"* que se desarrolla a su alrededor (Salais 2006). Los indicadores forman parte de una discusión abierta acerca de qué es una buena investigación, pero también tienden a adquirir vida propia, convertirse en objetivos estratégicos y fomentar ciertas motivaciones autónomas. *"People learn to manage the reporting of performance"* (Chapman 2006, 13). Dicho de otra manera: "el éxito en el proceso de evaluación puede convertirse en un objetivo más importante que el éxito en la investigación misma" (Brook 2002, 176). Las organizaciones y los individuos descubren que la reputación es construida por un determinado *rating* y *ranking*, al que terminan adaptando su comportamiento. Lo que tuvo su origen en el deseo de promover la consecución de los objetivos pretendidos por la investigación científica termina produciendo el indeseable resultado de una investigación que se incapacita para la producción de lo nuevo.

3. La producción científica de lo nuevo

Entre los principales efectos indeseados de la evaluación científica el que más debería inquietarnos es la posibilidad de que su uso irreflexivo contribuya a dificultar la producción científica de lo nuevo, que es el objetivo nuclear de la investigación científica. Puede ser que el aumento de la evaluación se rija por un principio de utilidad marginal decreciente, de manera que lo que comienza separando el saber de la insignificancia termina produciendo una estandarización, actuando a favor del *mainstream*.

Para entender a qué se debe este fenómeno es necesario reflexionar un momento sobre cómo funciona la novedad en el sistema científico. Por definición, un pensamiento que sea realmente nuevo tiene que suscitar resistencias o, simplemente, no puede ser reconocido como tal. En cambio, el sistema de la mutua evaluación tiende, por su propia

racionalidad, a dificultar la innovación. El incremento de las evaluaciones a todos los niveles promueve una "espiral de adaptación" en virtud de la vigilancia recíproca que dificulta la innovación a través del disenso. La investigación innovadora corre el peligro de ser poco valorada en la medida en que puede chocar contra el criterio de los evaluadores. La lógica de las citas normaliza la investigación en la medida en que entroniza lo "moderadamente nuevo", mientras que los innovadores más radicales —que generalmente no tienen pares que los citen— permanecen invisibles.

Un sistema en el que todos sus actores se orientan por los mismos indicadores corre el peligro de perder su diversidad y capacidad de innovación. Un procedimiento de evaluación que no sea consciente de este efecto secundario puede provocar que disminuya el apetito por el riesgo intelectual y que la investigación crítica deje de valer la pena.

Donde mejor se comprueba este dilema ante la novedad es en la "paradoja del proyecto de investigación": si uno dice que no sabe realmente lo que va a resultar al final, será rechazado por pedir una cantidad de dinero para no se sabe qué; pero si uno anticipa el resultado, entonces se hace acreedor del reproche de que carece de sentido financiar una investigación sobre lo que ya se sabe. Por eso los proyectos de investigación deben hacerse, por así decirlo, "medio bien", sin dejar demasiado abierto el resultado ni anticiparlo excesivamente: porque nuestros sistemas de evaluación sólo están capacitados para percibir, repito, lo "moderadamente nuevo".

La producción del conocimiento obedece hasta un cierto punto a dinámicas anárquicas y no es algo completamente gobernable. Lo único que cabe hacer —y no es poco— es establecer las condiciones que faciliten el surgimiento de nuevos saberes. Pero una cultura que favorezca la novedad y que esté abierta al futuro requiere dos cosas: renuncia al control y paciencia (Koschorke 2004, 156). La *longue durée* del trabajo científico necesita un clima de estabilidad y confianza, que no puede estar permanentemente interrumpido por las exigencias de rentabilidad a corto plazo.

Tal vez no tengamos otro remedio que reconocer que la evaluación, tan necesaria como es, no sirve más que para asegurar un nivel mínimo, para la *"normal science"* (Kuhn 1962), pero no para promover la verdadera innovación científica. Ahora bien, ¿y si fuera todavía peor y la evaluación actuara como un impedimento? ¿Habrían sobrevivido Planck, Frege, Keynes o Wittgenstein en nuestro actual sistema de evaluación? Probablemente hay que resignarse al hecho de que las ideas innovadoras no pueden ser reconocidas como tales por sus contemporáneos. Que los vagos se amparen en su genialidad no reconocida no es un motivo ni para suprimir las evaluaciones pero tampoco para confiar que ellas produzcan milagrosamente la excelencia.

4. La especificidad de los saberes humanísticos

El cuarto y último eje que advierte de las limitaciones del sistema de evaluación tiene que ver con su falsa universalidad, es decir, con el hecho de que muchos científicos impugnan la evaluación porque entienden que es un sistema que procede de una determinada cultura científica y no respeta la especificidad de otras posibles culturas.

La presión del mercado sobre la universidad lleva a que la investigación y la docencia se orienten más hacia la aplicación o los saberes inmediatamente útiles. Las ciencias humanas, sociales y jurídicas —al igual que una parte importante de las ciencias básicas— son saberes que tienen otra relación con los criterios de utilidad, en los que se cultiva un sentido crítico en relación con esa inmediatez con el que otras disciplinas mantienen una mayor familiaridad. Estas disciplinas tienen una praxis que no se adapta bien a los criterios normalizados de cientificidad. Defender esa especificidad contribuye no solo a enriquecer la pluralidad de las ciencias sino también la posibilidad de una integración equilibrada de los saberes.

Hay una sensación de que los saberes humanísticos experimentan una invasión de métodos, criterios y enfoques

que no hacen justicia a su especificidad. De hecho, se ha producido una cierta estandarización de las publicaciones, algo que puede tener efectos negativos sobre la creatividad en estos ámbitos de la ciencia. Lo que vale como estándar es el artículo corto, altamente especializado, en una revista anglosajona y del que son responsables varios autores en una secuencia minuciosamente establecida. Las humanidades no pueden ser valoradas con los mismos criterios de relevancia y productividad que las especialidades técnico-científicas. Pero en esto no hay que ver solamente un problema sino también una oportunidad: gracias a la cultura de la interpretación en la que se desenvuelven los estudios humanísticos, el saber no queda reducido a la mera acumulación de datos y la universidad no se pliega a los imperativos del rendimiento productivo inmediata como máxima prioridad.

Tenemos el caso del *Standing Commitee for the Humanities Citation Index,* que desde el año 2000 sostiene que el *Citation Index* del ISI es completamente inadecuado para estas disciplinas y no debería por tanto ser utilizado por las autoridades europeas, entre otras razones porque constituye una sobre-representación de revistas americanas, porque supone un menosprecio a la cultura de las disciplinas humanísticas (de la que forman parte los libros y las actas de congresos), y porque se concentra en las publicaciones recientes, mientras que en las humanidades y ciencias sociales la vigencia de las publicaciones no se rige por el criterio de actualidad reciente (Peyraube 2002). Como advertía recientemente Helga Nowotny, actual presidenta del *European Research Council,* en una conferencia con un título provocador "¿Cuánta evaluación soporta el sistema científico?", las valoraciones de la ciencia han de respetar los criterios al uso en el seno de cada específica comunidad científica, como puede ser el caso, por ejemplo, de los libros en las humanidades (Nowotny 2012, 14).

En humanidades y ciencias sociales es muy frecuente que el destinatario de la publicación sea un público nacional (pensemos en ciertos ámbitos de la filología, la historia, el derecho o la sociología). Buena parte de su investigación se dirige a un público no científico (funcionarios, jueces, público

culto en general), para los que debe haber otros canales de comunicación diferentes de las revistas especializadas y en idiomas distintos del inglés. Si esta y otras especificidades no son reflejadas en los criterios de valoración, habremos equivocado nuestras estrategias para medir la excelencia de la investigación científica en todas sus dimensiones.

Conclusión: reflexividad y pluralismo contra la incertidumbre

Me gustaría concluir estas notas con una constatación y dos recomendaciones. La primera, que las limitaciones comprobadas en nuestros sistemas de evaluación no deberían llevarnos a su impugnación general sino a reconocer que, como todos nuestros sistemas de medición, son inexactos, paradójicos y necesitados de interpretación. Los indicadores no son más que eso: indicadores. No están libres de toda duda y al margen de cualquier contexto. No sería una mala conclusión de todo esto si concluyéramos que la ciencia es algo más complejo de gobernar de lo que suelen pensar sus gestores menos reflexivos.

Esta constatación de que hay una dimensión de incertidumbre que ninguna cuantificación puede nunca eliminar invita a manejar nuestros sistemas de evaluación respetando dos principios fundamentales: la reflexividad y el pluralismo, que responden al procedimiento mediante el cual los sistemas inteligentes se enfrentan a la complejidad.

En primer lugar, se impone hacer un uso reflexivo y crítico de los indicadores y la evaluación. Del mismo modo que los datos no hacen superflua la interpretación, las evaluaciones no sustituyen completamente a las decisiones políticas. La tendencia a delegar en los evaluadores decisiones políticas (objetivos, criterios, equilibrios…) sobrecarga a los evaluadores. Las evaluaciones sirven para determinar en qué medida se han alcanzado los objetivos propuestos, pero no definen esos objetivos.

La cuestión de los criterios de evaluación debe ser objeto de una permanente discusión pública. La reflexión en torno a

la pertinencia de nuestros procedimientos de evaluación sólo puede tener lugar dentro de una discusión acerca de nuestros valores en relación con la actividad científica. Es lógico que en este debate haya cuestionamientos y críticas, que sirven para impulsar la reflexividad de la evaluación; una evaluación reflexiva es aquella que se conduce consciente de sus límites y paradojas, acompañada siempre por una interrogación acerca de sus posibles efectos y evaluando sistemáticamente sus propios criterios. Se trata de examinar críticamente y convertir en objeto de debate público los criterios por los cuales es auditada la investigación.

La cuestión fundamental que debe presidir esta reflexividad es saber si con el instrumento de la evaluación alcanzamos el fin de asegurar la calidad científica. O dicho de una manera provocativa: ¿podemos inventarnos indicadores de la calidad de la investigación que puedan darnos más de lo que esperamos de la ciencia, o sea, que nos proporcionen más innovación y menos "ciencia normal" asegurada?

El otro instrumento del que disponemos para facilitar el reconocimiento de la innovación científica es la pluralización de los criterios de calidad. Hay un amplio ámbito de investigación para cuya valoración no siempre disponemos de indicadores cuantitativos por lo que los pares no tienen más remedio que ponderar otros criterios de juicio. Los criterios de la comunicación de corto plazo entre especialistas son insuficientes para juzgar la validez en nuestras materias. Una posible solución es pluralizar los criterios en la línea sugerida por ciertos expertos, a la manera de radar que incluya cinco ámbitos: ciencia y conocimiento certificado; educación y formación; innovación y profesionales; política pública y cuestiones sociales; colaboración y visibilidad (Spaapen / Dijstelbloem 2007). Se trataría, por tanto, de pluralizar los criterios con los que juzgamos la calidad de la investigación en lugar de confiarlo todo a unos pocos valores suministrados por unos pocos especialistas.

Así pues, el problema de las listas, los indicadores y los *rankings* es que haya tan pocos. ¿Por qué no pensar que la solución al problema planteado está en diseñar una pluralidad

de valores que responda a la pluralidad de ciencias, perfiles y centros en los que se desarrolla esa "lucha de las facultades" que para Kant constituía el verdadero diálogo de los saberes?

BIBLIOGRAFÍA:

Bénabou, Roland / Tirole, Jean (2003), *"Intrinsic and extrinsic motivation",* en *Review of Economic Studies* 70 (3), 489-520.

Brook, Richard (2002), *"The Role of Evalution as a Tool for Innovation in Research",* Max Planck Forum 5, Innovative Structures in Basic Decision Research. Ringberg Symposium, 173-179.

Chapman, Chris (2006), *"Joining accountability and autonomy in reasearch",* en *Foresight Europe* 2 (march), 13-14.

Day, Patricia / Klein, Rudolf (1990), *Age of Inspection. Inspecting the Inspectors,* London: Rowntree Foundation.

Fehr, Ernst / Gächter, Simon (2002), *Do Incentive Contracts Crowd Out Voluntary Cooperation?,* Institute for Empirical Research in Economics, Working Paper, nº 34.

Greshoff, Rainer /Knerr, Georg / Schimank, Uwe (eds.) (2003), *Die Transintentionalität des Sozialen,* Wiesbaden: Westdeutscher Verlag.

Hornbostel, Stefan (1997), *Wissenschaftsindikatoren. Bewertungen in der Wissenschaft,* Opladen: Westdeutscher Verlag.

Koschorke, Albrecht (2004), *"Wissenschaftsbetrieb als Wissenschaftsvernichtung. Einführung in die Paradoxologie des deutschen Hochschulwesens",* en Kimmich, Dorothee / Thumfart, Alexander (eds.), *Universität ohne Zukunft?,* Frankfurt: Suhrkamp, 142-156,

Kuhn, Thomas (1962), *The Structure of Scientific Revolution,* University of Chicago Press.

Miller, Peter (2001), *"Governing by Numbers: Why Calculative Practices matter",* en *Social Research* 68/2, 379-396.

Neave, Guy (1988), *"On the cultivation of quality, efficiency and enterprise: An overview of recent trends in higher education in Western Europe, 1986-1988",* en *European Journal of Education* 23 (1-2), 7-23.

Peyraube, Alain (2002), *"Project for building a European citation index for humanities",* en *Reflections*, European Science Foundation, december, 14-15.

Porter, Theodore M. (1995), *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life,* Princeton University Press.

Power, Michael (1977), *The Audit Society: Rituals of Verification,* Oxford University Press.

Nowotny, Helga (2012), *Auf der Suche nach Exzelenz. Wie viel Evaluierung verträgt das Wissenschaftsystem?,* Göttingen: Wallstein.

Salais, Robert (2006), *"Reforming the European social model and the politics of indicators: from the unemployment rate to the employment rate in the European strategy",* en Maria Jepsen / Amparo Serrano (eds.), *Unrawpping the European Social Model,* Bristol: The Policy Press, 189-212.

Sitkin, Sim / Stickel, Darryl (1996), *"The Road to Hell: The Dynamics of Distrust in an Era of Quality",* en Roderick Kramer / Tom Tyler (eds.), *Trust in Organisations: Frentiers of Theory and Reaearch Thousand Oaks,* Cambridge: Sage Publications, 196-215.

Spaapen, Jack / Dijstelbloem, Huub (2007), *Evaluating Research in Context. A Method for Comprehensive Assessment,* Den Hag: Consultative Committee of Sector Councils for Research and Development

Tucci, Christopher (2006), *"Why Europe will never have accountability in research",* en *Foresight Europe* 2 (march), 27-29.

Weingart, Peter (2005), *Die Wissenschaft der Öffentlichkeit. Essays zum Verhältnis von Wissenschaft, Medien und Öffentlichkeit,* Weilerwist: Velbrück.