Reflexiones metodológicas sobre la evaluación académica

Juan Ignacio Piovani

Introducción[1]

En los últimos años se han intensificado notablemente las actividades de evaluación en los ámbitos académicos y científicos de Argentina. Me refiero, en particular, a la evaluación sistemática e institucionalizada que, evidentemente, ha acompañado un proceso mucho más amplio de profesionalización e institucionalización de las prácticas académicas –y, muy especialmente, las científicas–.

Este proceso, que ha tenido impactos directos en el sistema universitario y en el científico, en sus instituciones y en las trayectorias de los docentes-investigadores, con efectos ambivalentes, ha comenzado a ser objeto de análisis por parte de diverso tipo de actores,[2] y ha dado lugar a un debate cada vez

[1] Este artículo se basa en la presentación relazada en el marco de las Jornadas "Investigación y Evaluación en Humanidades y Ciencias Sociales", organizadas por la Secretaría de Investigación de la Facultad de Humanidades (UNLP) y por el Instituto de Investigaciones en Humanidades y Ciencias Sociales (UNLP/CONICET) en 2014. Esta versión revisada y ampliada será publicada en la revista *Política Universitaria.*

[2] Por ejemplo, en el ámbito de la Presidencia de la Agencia Nacional de Promoción Científica y Tecnológica se creó en 2014 la Comisión Asesora en temas de evaluación. Más específicamente, en el campo de las ciencias sociales se puede citar el trabajo de una comisión del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), que se materializó en la Resolución 2249/14: "Bases para la Categorización de Publicaciones Periódicas en Ciencias Sociales y Humanidades." El Consejo de Decanos de Facultades de Ciencias Sociales y Humanas (CODESOC), por su parte, ha publicado una declaración titulada "Criterios para la evaluación de las ciencias sociales y humanas, y la jerarquización de la investigación científica con impacto social." Y a partir de la iniciativa de profesores e investigadores de diversas universidades y cen-

más intenso y articulado en el marco del cual se sitúan las reflexiones que propongo en este artículo.

El objetivo que se persigue, sin embargo, no es primariamente analizar las consecuencias que ha teniendo la lógica de evaluación en las trayectorias individuales o en el sistema, sino reflexionar, desde el punto de vista metodológico, sobre los procesos de evaluación en sí y, concomitantemente, aunque de manera mucho más tangencial, sobre el modo en que se han llevado a la práctica en Argentina.

Este tipo de reflexión resulta relevante porque en la medida que se profesionaliza, la evaluación adquiere una dimensión burocrática, casi ritual. Consecuentemente, en el marco de procesos que se vuelven rutinarios, se pueden ir desdibujando los objetivos de la evaluación y distanciarse gradualmente de las preguntas fundamentales que la orientan y de los principios que la justifican desde el punto de vista político-institucional.

Por supuesto que no se trata de un problema que afecte exclusivamente a la evaluación, o que se restrinja a prácticas del mundo académico. No obstante, puede resultar ilustrativo señalar que esta misma situación ha sido materia de análisis crítico en relación con algunas de las prácticas que son objeto de la evaluación académica, como la docencia y la investigación. En este último caso, por ejemplo, la metodología critica ha denunciado la fetichización de la técnica y sus usos rituales y rutinarios, independientemente de los objetos en estudio y de las formas más apropiadas de abordarlos (Marradi, 1996).

Análogamente, en la evaluación es crucial preguntarse sobre cuáles son los instrumentos, las técnicas y los procedimientos más adecuados para evaluar qué tipo de aspectos y en qué contextos. El esfuerzo por no perder de vista los fundamentos de la evaluación podría llevarnos a plantear preguntas incluso más básicas –que no pretendo de ninguna manera responder de manera exhaustiva–: ¿Qué significa evaluar? ¿Qué se evalúa? ¿Quién evalúa? ¿Para qué se evalúa? ¿Cómo se evalúa?

En la discusión que propongo puede ser útil diferenciar, al menos analíticamente, tres planos o niveles, aún cuando en la práctica de la evaluación sus

tros de investigación se conformó la Comisión Interinstitucional de elaboración de criterios de evaluación para las humanidades y ciencias sociales (CIECEHCS), que ha tenido una labor muy intensa en la promoción de debates, la organización de jornadas de discusión y la publicación de documentos específicos.

confines no sean siempre claros y tajantes. Como se verá más adelante, estos niveles encuentran un cierto correlato en lo que la metodología cuantitativa define como "proceso de conceptualización y operacionalización", que implica reconocer que la investigación científica (o por analogía la evaluación) no aborda su objeto –complejo y no directamente observable– sino a través de propiedades conceptuales para cuya traducción operativa se requerirá la selección de una pluralidad de indicadores (Lazarsfeld, 1973). También pueden entenderse los niveles propuestos a partir de la consideración de una operación intelectual clave en los procesos de evaluación: la clasificación. En este sentido, se puede identificar un primer nivel más general (clasificación$_a$), que involucra la tarea de descomponer analíticamente la extensión un concepto (en este caso aquél que define al objeto/sujeto de la evaluación o a cada una de sus dimensiones de interés) en clases a partir de un *fundamentum divisionis*[3], un segundo nivel (clasificación$_b$) que alude al resultado estructural de una clasificacióna, es decir, un listado tangible de clases o esquema clasificatorio, y un tercer nivel (clasificación$_c$) que implica la asignación de objetos concretos a las clases o tipos ya establecidos a través de la clasificación$_a$ y materializados en la clasificación (Marradi, 1990).

Teniendo en cuenta estas consideraciones, y volviendo al caso de la evaluación, es posible señalar una serie de aspectos problemáticos y preguntas específicas ─seguramente no todas─ que permiten definir cada nivel:

1. En primer lugar, el nivel conceptual: ¿Qué criterios sustentarán los procesos de evaluación? ¿Qué propiedades y dimensiones de los fenómenos y de los sujetos evaluados serán tenidas en cuenta? ¿Qué significan? ¿Cómo se definen conceptualmente y con qué términos las expresamos? ¿Quién decide o participa de su definición? ¿Son fijas en el tiempo y universales en su alcance? ¿Cómo se relacionan con los diferentes objetos que se evalúan?

2. En un segundo nivel encontramos el problema más clásicamente metodológico, que tiene que ver con la operacionalización de los criterios, las propiedades conceptuales y las dimensiones señaladas precedentemente:

[3] Esto también contempla las divisiones de la extensión efectuadas en simultáneo sobre la base de varios fundamentos (tipologías), y aquellas operadas en cadena sobre conceptos de generalidad decreciente (taxonomías).

¿Qué referentes o indicadores nos permitirán conocer el estado de un sujeto u objeto x con respecto a cada una de las dimensiones de interés? ¿Cómo los definiremos operativamente? ¿Se les puede atribuir a todos el mismo peso relativo? ¿O requieren de algún tipo de ponderación? ¿Qué operaciones intelectuales exige la aplicación de las definiciones operativas para la evaluación de casos concretos? ¿Se pueden reducir exclusivamente a la medición? El resultado de este trabajo se materializa habitualmente en un producto tangible: las tan mentadas grillas de evaluación.

3. Finalmente, un tercer nivel, de carácter más procedimental, se relaciona con la aplicación de las grillas de evaluación y, como resultado de ello, con la asignación de puntajes o con la atribución de un caso a una clase determinada. Esta tarea, además, habilita el proceso de análisis de los resultados, que a su vez justifica la toma de decisiones, sobre la base de la evaluación, en el ámbito institucional responsable de gestionarla.

Con este esquema tripartito como punto de referencia, a continuación presentaré algunas cuestiones problemáticas relacionadas con cada uno de los niveles, para luego retomarlas, a manera de cierre, vinculándolas con algunas observaciones críticas acerca de los procesos de evaluación implementados en diversos ámbitos académicos de Argentina.

Los niveles de la evaluación

El nivel conceptual

La expansión de la cultura de la evaluación implica que ésta alcance cosas muy diversas: desde instituciones –universidades, centros especializados– a trayectorias de investigadores, pasando por carreras de grado y posgrado, programas y proyectos de investigación, prácticas docentes, publicaciones, entre otras. No se pueden desatender, por lo tanto, las preguntas más elementales sobre qué es lo que se evalúa, cómo se lo define y conceptualiza, qué lo caracteriza. Y esto, a su vez, exige considerar, por un lado, el "objeto" de la evaluación (por ejemplo: cierto tipo de prácticas –de investigación, docencia, etcétera– y sus productos o resultados) y, por el otro, el "sujeto" evaluado (institución, carrera, investigador/a, docente, etc. ya sea en tanto ámbito en el cual las prácticas evaluadas se desarrollan o en cuanto sujeto que las lleva a cabo).

Pero también hay que tener en cuenta la importancia de considerar lo que hemos denominado "objetos" y "sujetos" de la evaluación de manera articulada,[4] ya que los objetos no se producen en el vacío, sino que son, en general, prácticas de sujetos enmarcados institucionalmente y que se atienen, en mayor o menor medida, a un conjunto de reglas más o menos formalizadas y explicitadas.

Al pensar en este par imbricado "objeto-sujeto", se pueden poner en evidencia algunos aspectos importantes a la hora de definir criterios y de seleccionar dimensiones para la evaluación. En primer lugar, hay una cuestión de carácter político institucional para nada menor: quién o quiénes, y cómo, definirán dichos criterios y seleccionarán las propiedades y dimensiones a ser consideradas. ¿Se trata de una competencia de los funcionarios técnicos de las instituciones responsables de gestionar la evaluación? ¿O debe ser materia exclusiva de los académicos? Y en este caso, ¿qué perfiles serían los más adecuados? ¿Quién los selecciona y a través de qué mecanismos? ¿La experticia en docencia y/o investigación, asegura por sí sola los conocimientos necesarios para intervenir en el diseño de la evaluación? En todo caso, y esto es una demanda recurrente por parte de los sujetos evaluados, los procedimientos para seleccionar a quienes participan de las decisiones que subyacen a la evaluación, así como los criterios de evaluación mismos, deben ser lo más públicos y transparentes posibles.

En segundo lugar, e independientemente de los procesos institucionales que se establezcan para decidir los criterios generales de la evaluación, ha de tenerse en cuenta su carácter multidimensional. Es evidente que la complejidad de las prácticas académicas y de las instituciones en las que se desarrollan exige enfoques de evaluación que consideren una multiplicidad de dimensiones con el fin de evitar reduccionismos y sesgos. Se trata de una cuestión ampliamente reconocida en la investigación científica, materializada en textos de metodología al menos desde los años 50, cuando ya se proponían formas de abordar objetos complejos desde diversos puntos de vista con el fin de reconstruirlos de modo más completo y ajustado.

[4] La combinación de ambos aspectos –objeto/sujeto– produce una amplia gama de situaciones particulares que pueden requerir atención específica. Tomemos por caso lo siguiente: al evaluar el objeto prácticas y resultados de investigación puede ser importante considerar algunas dimensiones del sujeto como la edad o el contexto institucional de inserción para poder definir el conjunto de criterios adecuados, porque el peso relativo de un artículo de revista o del dictado de un seminario de posgrado, por ejemplo, no es idéntico en todas las instancias de una trayectoria académica o para un perfil docente o de investigador.

En tercer lugar, la experiencia histórica muestra que los criterios no pueden ser fijos y, por lo tanto, requieren una revisión permanente. Piénsese, por ejemplo, en el caso del título de doctor: en la actualidad existe cierto consenso en torno de su importancia para valorar una trayectoria al momento de decidir un concurso docente o el ingreso a la carrera de investigador. Pero tan sólo 15 o 20 años atrás su peso relativo en la evaluación académica era muy diferente. Y ya hay cierta evidencia, especialmente si establecemos comparaciones con otros países de la región, como Brasil, de que los posdoctorados comienzan a ser considerados cada vez más como un requisito de acreditación para desempeñar ciertas funciones para las cuales, hasta ahora, valía típicamente el doctorado, o incluso la maestría.

Otro aspecto importante a señalar es que el conjunto de criterios que se formule, así como las propiedades y dimensiones que se tengan en cuenta para la evaluación, tienen ─en cierto sentido─ un carácter normativo. Es decir, contribuyen a instituir un modelo de perfiles y de prácticas académicas que se toma como parámetro y como patrón de comparación al evaluar casos empíricos concretos. Por otra parte, y esto ya ha sido objeto de investigación, estos modelos normativos tienen performatividad, en la medida que orientan las prácticas y promueven ciertas decisiones estratégicas, en detrimento de otras, sobre qué priorizar en el marco del desarrollo de una carrera académica (véase al respecto Fernández Lamarra & Marquina, 2012) o de la definición del perfil de un posgrado o de una institución.

En efecto, en nuestra práctica cotidiana es cada vez más frecuente encontrar comentarios acerca de las nuevas actitudes de los estudiantes de grado en torno de los promedios de calificaciones, sabedores de su importancia a la hora de asignar becas; o de investigadores jóvenes que planifican sus publicaciones a partir de un análisis de los potenciales espacios de publicación considerados de Nivel 1 por parte de las instituciones científicas, por citar solo un par de ejemplos. Por lo tanto, al establecer conjuntos de criterios es importante mirar más allá del acto de evaluación específico y considerarlo como parte de la política científica y de educación superior, con efectos que trascienden ampliamente la evaluación en sí.

Por otra parte, conviene recordar que la definición normativa *a priori* de criterios y de dimensiones presenta muchas limitaciones y, por lo tanto, podría ser beneficioso tener en cuenta, al menos complementariamente, una

perspectiva descriptiva que también aporte a la construcción de los perfiles a ser evaluados. Esto implica utilizar como recurso los resultados de la investigación empírica sobre los objetos y sujetos de la evaluación (en qué consisten sus prácticas; cómo las definen, reconocen y jerarquizan los propios actores, etcétera), por un lado, y sobre los procesos de evaluación mismos, por el otro.

Finalmente, otro aspecto desafiante para la evaluación, desde el punto de vista metodológico, es que habitualmente los criterios, las propiedades conceptuales y sus dimensiones tienden a enfocar aspectos "estáticos". En este sentido, se priorizan los productos o resultados tangibles que, por su materialidad (por ejemplo, un proyecto escrito, un libro o un artículo) resultan más asibles. Pero esos productos se inscriben en procesos "dinámicos" que también puede ser importante considerar en la evaluación y que presentan retos particulares para su definición y su aprehensión cognoscitiva.

El nivel de los indicadores

La definición de los criterios en los que se sustenta la evaluación, y de las propiedades conceptuales y las dimensiones a ser consideradas a tal fin, es una cuestión ineludible que condiciona todo el proceso de evaluación. Sin embargo, resulta estéril a los efectos prácticos si no se establecen operaciones de mediación ya que, como se indicó más arriba, por su complejidad y nivel de abstracción las propiedades conceptuales y sus dimensiones no resultan directamente observables.

Este es un problema que en la tradición canónica de la investigación social ha sido abordado profusamente. El clásico texto de Lazarsfeld (1973), "De los conceptos a los índices empíricos", constituye un punto de referencia obligado. Expresado de manera simple, su argumento es que los fenómenos de interés sólo pueden ser conocidos a través de referentes empíricos observables que deben estar semánticamente vinculados con alguna de las múltiples dimensiones en cuestión ─por eso indicadores─ y, a través de ellas, con las propiedades conceptuales que en el nivel teórico más abstracto definen, en nuestro caso, el núcleo de evaluación.

Esta esquemática presentación introductoria deja en evidencia, una vez más, el problema de la multidimensionalidad. Ya se había señalado su relevancia a la hora conceptualizar los objetos y sujetos de evaluación, por la complejidad y diversidad de actores, prácticas y productos del trabajo acadé-

mico; pero también resulta fundamental a la hora de seleccionar los referentes que permiten, en el proceso de evaluación, explorar y determinar el estado de los casos evaluados con respecto a cada una de ellas. Es en esta línea que Lazarsfeld (1973) aboga por la construcción de índices empíricos, que no son más que el producto de la combinación de indicadores.

El uso de varios indicadores en simultáneo para recomponer una dimensión o propiedad conceptual compleja es en la actualidad una práctica estándar, con la que también se suele relacionar la articulación de varias mediciones o registros independientes de cada uno de los indicadores, con el fin de obtener resultados más estables y confiables. En la literatura metodológica de los años 50 ya aparece tematizada esta propuesta, que se nombra de diversas maneras: operacionismo múltiple, operacionismo convergente, validación convergente, entre otras. En los años 60 se usó por primera vez el término triangulación[5] para referirse a esta estrategia: "…una vez que una proposición ha sido confirmada por dos o más mediciones independientes, la ambigüedad en cuanto a su interpretación se reduce significativamente. La evidencia más persuasiva se da a través de la triangulación en el proceso de medición" (Webb et al., 1966). Y este mismo enfoque era práctica común, mucho tiempo antes, en otros campos científicos, comenzando por la Astronomía del siglo XVIII (Piovani, 2008).

Volviendo a la selección de indicadores, un aspecto clave es valorar la validez, entendida desde el punto de vista metodológico como una propiedad de la relación entre el indicador y un concepto más general. Se ha sostenido, al respecto, que tal la relación no es unívoca, y que el indicador siempre ostenta una parte indicante, que alude al contenido semántico compartido con el concepto en cuestión –y que por lo tanto lo justifica como indicador–, y una parte extraña que, en todo caso, comparte contenido semántico con otro/s concepto/s. Como no existen procedimientos "objetivos" para determinar la validez de un indicador, su selección deberá basarse en un cuidadoso examen que tenga en consideración aspectos teóricos sustantivos.

También se ha tematizado, respecto de los indicadores, el problema de su supuesta universalidad y equivalencia. Según muchos autores, ellos pueden

[5] En la metodología contemporánea el término triangulación es polisémico: se refiere no sólo a la combinación de mediciones o de indicadores, sino a también a distintas formas de articulación entre teorías y/o métodos, o al trabajo conjunto de varios investigadores sobre un mismo tema, etcétera.

tener una especificidad situacional tal que los vuelve –potencialmente– referentes empíricos de fenómenos diversos en contextos diferentes, poniendo en duda su universalidad y equivalencia (Bynner & Chisholm, 1998; Fideli, 1998). Esto llama la atención sobre la necesidad de definir –al menos en alguna medida– no sólo criterios, estándares y dimensiones específicas, sino también indicadores particulares de acuerdo con los perfiles de las prácticas y de los sujetos a ser evaluados, considerando los contextos históricos e institucionales. Además, alerta sobre la traslación mecánica y acrítica de esquemas de evaluación de un ámbito a otro.

Otra cuestión relevante es la ponderación, que adquiere significación cuando se reconoce que, aún siendo pertinentes, diferentes indicadores no tienen el mismo peso relativo para reconstruir la complejidad de una práctica o de una trayectoria académica. La ponderación es moneda corriente en la evaluación: por ejemplo, al considerar la producción de un docente-investigador suele asignársele un peso mayor a un artículo de una revista con referato que a una ponencia de un congreso. Sin embargo, los esquemas de ponderación tienen que ser el resultado de análisis muy minuciosos y deben que estar sujetos a revisiones. Asimismo, pueden variar según los perfiles evaluados, de manera tal que un mismo indicador adquiere pesos relativos diferentes para cada uno de ellos. Esto último resulta obvio en el caso de la evaluación de investigadores: el promedio de calificaciones de la carrera de grado, por ejemplo, tiene mucho mayor valor en edades tempranas –al decidir la asignación de una beca– pero se va desdibujando a medida que se avanza en la carrera.

Una selección pertinente de indicadores, aun siendo fundamental, no resuelve todos los problemas operativos de la evaluación, ya que para poder determinar el estado de un caso *x* en un indicador *y* debemos contar con una definición operativa, es decir, un conjunto de reglas y convenciones que hagan posible su "medición". No obstante, a propósito de la medición, es oportuno recordar que las definiciones operativas pueden exigir, como parte de sus reglas y convenciones, la puesta en práctica de variadas operaciones intelectuales, como la clasificación o el conteo, además de la medición.

Aunque a primera vista parezca un detalle menor, la afirmación precedente nos da la oportunidad de poner en evidencia la tendencia –también en las prácticas evaluativas– a sobrevalorar la medición o a considerar medición cualquier acto de atribución de números –por más arbitraria que sea– al estado

de una unidad de observación en un indicador determinado. Estas prácticas "abusivas" de la medición son multifacéticas: por un lado implican el mero estiramiento semántico del término, en línea con la perspectiva hegemónica consagrada en el clásico artículo de Stevens (1946) sobre los niveles de medición, que llega incluso al contrasentido de proponer la "medición nominal". Por otro lado, aluden a la tendencia a privilegiar los atributos genuinamente mensurables (aunque sean muy limitados) para dar cuenta del objeto de evaluación o a forzar su conversión en tales, habilitando incluso cálculos aritméticos completamente injustificados a partir de números que no establecen ninguna propiedad de magnitud, y que mucho menos implican distancias iguales entre puntos adyacentes de una escala, sino que son meros rótulos arbitrarios empleados convencionalmente en reemplazo de una categoría cualitativa.

La discusión en torno de las definiciones operativas conduce, además, a otro aspecto muy significativo: la fiabilidad. Con este término nos referimos a la capacidad de representar el estado "real" de una unidad según las convenciones de la definición operativa. Este problema fue reconocido tempranamente en la investigación social, cuando Ebbinghaus, un destacado psicólogo alemán inscrito en la tradición psicofísica wundtiana, criticó la atribución de valores a ciertos atributos intelectuales de los niños a partir de las meras declaraciones subjetivas de sus docentes, como era habitual en los experimentos psicológicos de la época, y llamó la atención acerca de la necesidad de contar con mecanismos de medición más estables y consistentes (Piovani, 2006).

Finalmente, ha de considerarse una cuestión tal vez más sustantiva, que remite a la posibilidad –o no– de encuadrar todas las dimensiones relevantes de la evaluación en un esquema puramente cuantitativo. Si bien es cierto que la mayoría de las grillas de evaluación en uso tiene un componente cualitativo, hay quienes consideran que el enfoque cualitativo debería ser empleado integralmente, y no de manera residual. Esto pone sobre el tapete otra discusión, que nos lleva a considerar el último de los niveles propuestos –el procedimental/operativo– y que alude a las formas concretas en que se realiza la evaluación de un caso. Por supuesto que el problema procedimental (o de práctica de la evaluación), como veremos a continuación, no se limita a las estrategias cualitativas en las que las competencias del evaluador adquieren una evidente centralidad, sino que también tiene una relevancia fundamental en los casos de evaluaciones que se guían por grillas cuantitativas pre-codifi-

cadas, con altos niveles de detalle y desagregación.

El nivel procedimental

En la práctica de la evaluación, una situación conflictiva bastante habitual deriva del hecho de que los evaluadores, que en general también son docentes y/o investigadores (y que pueden desempeñar –o haber desempeñado– múltiples funciones institucionales) tienen sus propias ideas sobre los más diversos temas académicos y, en consecuencia, pueden poner en juego una suerte de *curriculum* oculto (adaptando la categoría ampliamente conocida en el campo educativo) o una agenda propia que entra en contradicción con los criterios de evaluación o las grillas en cuya elaboración generalmente no han participado.

Por lo tanto, aun contando con una grilla de evaluación detallada y consensuada, su aplicación no es lineal ni mecánica. Por el contrario, se trata de un proceso en el que pueden aparecer variados inconvenientes. Y como se ha señalado en el apartado precedente, los problemas de este tipo no se limitan al componente cualitativo de las grillas estandarizadas –o a la evaluación cualitativa en general–, que sus detractores tienden a cuestionar por su "evidente" connotación subjetiva.

Recurriendo una vez más a la analogía con la investigación social, podemos ver que las limitaciones también se presentan en el caso de las estrategias que recurren a instrumentos estandarizados. En este campo, los problemas de administración del instrumento suelen englobarse bajo el rótulo de "errores no de muestreo". Pueden consistir en un uso inadecuado del instrumento, por desconocimiento de las reglas y convenciones que establecen las definiciones operativas para la asignación de valores al estado de una unidad en un indicador, o en simples errores de registro de información. Dadas las dificultades para cuantificarlos, gran parte de la literatura metodológica tendió a minimizarlos y se concentró, en cambio, en el cálculo de los errores de muestreo. Pero los textos metodológicos más actuales también le asignan mucha importancia a los errores "no de muestreo" y, basándose en investigaciones empíricas específicas, señalan que ellos pueden afectar severamente la calidad de los resultados de una investigación (y, por extensión, los de una evaluación). Este problema se torna particularmente serio cuando se reflexiona sobre sus posibles consecuencias, por ejemplo, en las trayectorias de las personas: como es sabido, un simple error de asignación de puntaje podría

decidir el acceso a un cargo docente universitario, o el ingreso a la carrera del investigador de CONICET.

Por otra parte, en el plano procedimental, un desafío central consiste en imaginar estrategias de evaluación que permitan controlar la subjetividad inherente a todo proceso de evaluación, para evitar que ésta derive en discrecionalidad o arbitrariedad. Con esta finalidad ha sido recurrente la idea de evaluación intersubjetiva, que se materializa en la forma de plenarios u otros mecanismos de evaluación colectiva. Al respecto, cabe señalar que estas estrategias tienen sentido bajo ciertas condiciones: que haya diversidad de perspectivas y que las diferentes voces estén en condiciones materiales de ser escuchadas y de influir en el proceso de evaluación.

Recurriendo nuevamente a una analogía con la investigación científica, podemos pensar –a la manera de referente– en el análisis de contenido clásico, que se basa en la construcción y aplicación de un marco o esquema de codificación a un *corpus* dado de materiales cualitativos (Krippendorf, 1980). En estas estrategias, para evitar los efectos arbitrarios de la autorrealización en la construcción de los esquemas, se suele recomendar que investigadores con diferentes formaciones e intereses elaboren marcos de codificación que luego se confrontan. De esta manera, se apunta a alcanzar una versión consensuada. Asimismo, se recurre a esta modalidad de confrontación para determinar el grado de acuerdo en el uso concreto del esquema de codificación al momento de analizar el *corpus.* De este tipo de prácticas de investigación se pueden derivar importantes enseñanzas para la evaluación intersubjetiva. Pero ellas también ponen en evidencia la necesidad de capacitar y entrenar a los evaluadores en las tareas específicas que esta práctica conlleva. Resulta bastante razonable pensar que algunos de los problemas procedimentales de la evaluación podrían minimizarse si se dedicaran ciertos esfuerzos a generar saberes compartidos en torno de la evaluación, de sus criterios subyacentes, de las dimensiones relevantes y de los mecanismos de asignación y control de puntajes, en caso de que los hubiera.

Por otra parte, es posible diseñar estrategias complementarias y de fácil ejecución para reforzar los intercambios de opiniones y pareceres que suelen caracterizar a los plenarios. Por ejemplo, cuando se trate de evaluaciones basadas en propiedades cuantitativas, o que resulten en una calificación numérica general, es posible contrastar las medias aritméticas y desvíos estándar

calculados a partir de los puntajes asignados por los diferentes evaluadores involucrados en una misma comisión. Diferencias considerables en estas medidas estadísticas podrían deberse a un efecto del azar en la conformación de los grupos evaluados o a inconsistencias en las asignaciones de puntaje. En el primer caso, los evaluadores trabajaron efectivamente sobre muestras desiguales (por ejemplo: un conjunto de proyectos, investigadores o artículos cualitativamente mejores o peores que el otro) y los puntajes divergentes simplemente reflejan esta situación. Pero las diferencias también podrían ser el resultado de criterios disimiles en la valoración de la "calidad" y la asignación de puntajes: un evaluador, por ejemplo, podría asignarle un 10 a un caso que considera muy bueno, mientras otro califica con 8 a un caso similar.

Más allá de las razones subyacentes, diferencias importantes en las asignaciones de puntajes exigirán una revisión integral de las evaluaciones. Claro que esto podría solucionarse desde un principio si todos los casos fuesen evaluados independientemente por varios evaluadores. Lamentablemente, la carga de trabajo y el volumen de casos a evaluar hacen que este ideal se vuelva –al menos muchas veces– impracticable.

Finalmente, parece oportuno señalar que, en la medida que se reconocen los complejos problemas procedimentales en materia de evaluación, y sus potenciales consecuencias arbitrarias –aún cuando se deban a errores involuntarios–, es fundamental que toda evaluación pueda estar sujeta a revisión, más allá del legítimo rechazo que genera en muchos ámbitos institucionales la creciente "cultura de la impugnación", que tiende a poner en duda todos los procesos de evaluación (desde los concursos docentes hasta las evaluaciones institucionales, pasando por la valoración de carreras de grado y posgrado, la asignación de becas, los ingresos a la carrera del investigador y sus promociones), así como a los actores involucrados en la evaluación (que en la mayoría de los casos son pares) y a las instituciones responsables de su gestión (Universidades, CONEAU, CONICET, FONCYT, etcétera).

Comentarios finales

A la luz de las consideraciones metodológicas precedentes, y atendiendo a algunos de los problemas señalados, a continuación presentaré una serie de observaciones y comentarios críticos acerca de los procesos de evaluación en diversos ámbitos académicos, basados en la experiencia en gestión y en

la participación en comisiones de evaluación, y enmarcados en algunos de los debates actualmente en curso sobre el tema. En efecto, la participación en comisiones de evaluación, por un lado, y en debates sobre evaluación organizados por diversos actores, por el otro, así como la lectura de algunos de los documentos que se han elaborado[6] en esos marcos, permiten recuperar diferentes argumentaciones en circulación, y relacionarlas con las discusiones metodológicas propuestas.

1. Al considerar los criterios de evaluación, se verifica una tendencia a rechazar los modelos vigentes –especialmente en el campo de la investigación– sobre la base de una supuesta oposición entre ciencias naturales y ciencias sociales. No obstante, en estas últimas se han encontrado significativos inconvenientes para producir un conjunto de criterios consensuados que reflejen la diversidad de prácticas y de estilos de producción y que, por ende, satisfagan a los diversos actores involucrados en las actividades académicas. Más compleja aún ha resultado la elaboración una grilla de evaluación única que logre condensar todas las propuestas y perspectivas.

2. El énfasis que se pone en pensar criterios completamente diferentes para las ciencias naturales, por un lado, y para las ciencias sociales, por otro, aun teniendo razones atendibles, puede obstaculizar el desarrollo de modos más creativos y comprehensivos de evaluación. En parte, porque se basa en la idea errónea de que las ciencias naturales y las ciencias sociales y humanas constituyen dos espacios monolíticos e inconmensurables. Aquel que suele tomarse como modelo de evaluación consensuado, producido en (o para) el ámbito de las ciencias naturales, y percibido como una imposición externa en las ciencias sociales, también afronta severas críticas –muchas de ellas en términos similares a los de los científicos sociales– por parte de docentes e investigadores de las ciencias naturales.

3. Lo anterior no anula las genuinas incomodidades que generan, en las ciencias sociales, muchos de los esquemas de evaluación vigentes. Sin embargo, una mirada alternativa permite poner en evidencia que las limitaciones

[6] Por ejemplo, los ya citados en la nota 2.

no derivan –en su totalidad– de la imposición del supuesto modelo evaluativo de las ciencias naturales, sino de esquemas que tienden a sobrevalorar las prácticas de investigación y, en sentido más estricto, aquellas que se ajustan a la investigación científica definida en términos canónicos. Como el peso relativo de dichas prácticas es mayor en las ciencias naturales, se concluye entonces que se trata de un modelo pensado exclusivamente para esas disciplinas. Pero en este punto conviene hacer dos aclaraciones. La primera es que también en el campo de las ciencias sociales y humanas existen tradiciones que se ajustan a dichos modelos. Piénsese por ejemplo en la Psicología experimental y en la Arqueología, o incluso en lo que Buroway (2005) denomina "sociología [académica] profesional"[7], en particular en su variante cuantitativa. La segunda aclaración es que, más allá de las reconocidas diferencias entra las diversas ciencias, este modelo es inadecuado para evaluar otras prácticas y productos de investigación no estándar –habitualmente más frecuentes en las artes, las ciencias humanas y algunas de las ciencias sociales– y, en general, para dar cuenta del conjunto de actividades académicas y actores que se desempeñan en el ámbito universitario, incluso en el de las ciencias naturales y las ingenierías. Si pensamos a la universidad desde una perspectiva más integral, se podrá reconocer que en ella se hacen muchas cosas, además de investigación científica en sentido estricto, e intervienen diversos actores, aparte de los científicos. Estas otras prácticas y actores también deberían ser pasibles de evaluación, aunque a partir de criterios y lógicas que sean adecuadas al objeto/sujeto y que eviten la ya criticada traslación mecánica de grillas pensadas para prácticas muy diferentes. En el caso argentino, esto implica tomar con cautela los intentos de llevar a la universidad los tipos de evaluación que se encuentran relativamente consolidados en instituciones más específicas, como el CONICET, y que tienen una larga tradición en evaluación. Esta cautela no implica, sin embargo, que muchos aspectos no puedan ser tenidos en cuenta como marco de referencia para el espacio universitario.

4. Pensando la evaluación en contextos académicos periféricos, como los

[7] Con esta expresión se refiere a un tipo de sociología que se desarrolla de modo relativamente autorreferente en el ámbito académico, que formula problemas de investigación específicos cuyo resultado tangible es un tipo de producto especializado ─el *paper*─ dirigido exclusivamente a un público de pares.

han llamado algunos autores, hay que tener en cuenta también el problema que Hanafi (2011) denomina "isomorfismo institucional", que consiste en tomar como modelo y parámetro para la evaluación –especialmente en la universidad– aquello que (se supone que) se hace en las instituciones de élite de los países centrales, en particular en Estados Unidos y Reino Unido. Esta perspectiva de la evaluación no tiene debidamente en cuenta, entre otras especificidades, las condiciones de producción y de trabajo, ni la importancia de la disponibilidad de recursos.[8] En contraste, se ha argumentado a favor de un tipo de evaluación que se ajuste a las particularidades nacionales e institucionales. Sin embargo, la sensibilidad contextual no puede ser una excusa para sostener el nacionalismo académico, prescindiendo de todo lo que se hace –y cómo se lo hace– en otros contextos. En otras palabras, el rechazo de una uniformidad transnacional y transcontextual absoluta tampoco puede ser la base propositiva de una evaluación a la medida de cada institución o de cada individuo, si es que se los considera partes de un sistema más amplio.

5. En relación con lo anterior, y teniendo en cuenta la experiencia histórica en evaluación, así como la diversidad de prácticas y de actores que constituyen su "objeto/sujeto", podría resultar razonable renunciar a la pretensión de contar con un único conjunto de criterios y una única grilla de evaluación, más allá de las diferencias entre ciencias sociales y naturales, o entre contextos nacionales e institucionales. En este sentido, es particularmente interesante revisar una propuesta de evaluación desarrollada en la Universidad de Helsinki, en la que participó Hebe Vessuri, y que ella misma comentó en una jornada de discusión sobre criterios de evaluación organizada en 2014 por la CIECEHCS y el IDES.[9] En dicha ocasión se explicó que los responsables de la evaluación, de manera dialógica con diversos actores institucionales, produjeron un consenso en torno de cinco modelos de evaluación para per-

[8] Tómese como ejemplo lo siguiente: la Universidad de Harvard, unánimemente considerada una institución de élite a nivel mundial, contó en el año 2014 con un presupuesto de 4,4 billones de dólares y un *endowment* de 36,4 billones de dólares (y tenía entonces cerca de 20.000 estudiantes y poco más de 4.500 docentes/investigadores). La Universidad de Buenos Aires, por su parte, contó para el mismo año con un presupuesto aproximado de 0,64 billones de dólares (y tenía más de 300.000 estudiantes y cerca de 30.000 docentes/investigadores).

[9] La relatoría de esta Jornadas se encuentra disponible en: http://cas.ides.org.ar/files/2014/08/Relatoria-Reunion-con-Hebe-Vessuri-IDES-CIECEHCS-120814.pdf

files específicos. Por supuesto que para cada perfil se establecieron reglas, parámetros y "criterios de calidad", pero se evitó la traslación mecánica de aquellos pensados para evaluar investigaciones de alcance internacional, por ejemplo, a otras cuya relevancia consistía en la transferencia de saberes para la solución práctica de problemas locales.

6. Con frecuencia, la crítica sobre las lógicas y los procesos de evaluación en curso va acompañada de propuestas de cambio bastante radicales. En algunas ocasiones, esto se basa en el supuesto –a mi juicio ingenuo, o ilusorio– de que es posible racionalizar, objetivar y anticipar todos los problemas involucrados en la evaluación, y ponerlos en juego en el diseño de un conjunto de dispositivos operativos definitivo. En cambio, podría ser útil considerar ajustes incrementales basados en la experiencia acumulada y en las críticas que se derivan del análisis permanente y constante de lo que se va haciendo en materia de evaluación. En este sentido, hay que tener en cuenta que los problemas de evaluación siempre tienen soluciones relativamente precarias, que exigen re-examen: identificar un inconveniente, y solucionarlo, puede casi inmediatamente generar otro problema para la evaluación, en la medida que ésta tienen efectos concretos sobre las trayectorias de las personas y sobre el sistema de educación superior y de ciencia y tecnología.

7. En los discursos sobre la evaluación que circulan actualmente también se encuentra un reclamo, cada vez más fuerte, por renunciar a la pretensión de cuantificar todos los aspectos relacionados con la evaluación de las prácticas académicas. En muchos casos, tanto en las ciencias sociales como en las naturales, este reclamo no se limita a la introducción de componentes cualitativos complementarios en las grillas predominantemente cuantitativas, sino que implica una apuesta por evaluaciones cualitativas integrales. Es frecuente encontrar que esta posición se basa en observaciones muy agudas sobre las concepciones de evaluación que subyacen a los modelos cuantitativos. Al igual que en la crítica a la investigación cuantitativa en general, se señala el riesgo de que este tipo de evaluaciones produzca resultados superficiales, incluso artificiosos, que no logren captar con suficiente profundidad lo que se evalúa. Pero en relación con esto se presenta una paradoja: en general, se trata de perspectivas con un desbalance entre instrumentos conceptuales y opera-

tivos. Esto significa que en el plano teórico existen valiosos aportes críticos sobre la evaluación; pero, en el plano operativo, (aun) se depende en gran medida de las conocimientos tácitos y personales de los evaluadores, porque se cuenta con un menor cúmulo de conocimientos técnicos e impersonales que garanticen cierta homogeneidad entre evaluadores y replicabilidad de criterios en el tratamiento de los casos evaluados. En otras palabras, hemos estado en condiciones de proponer discursos muy articulados sobre cómo mejorar la evaluación desde un punto de vista cualitativo, pero no siempre hemos conseguido traducirlos en esquemas de evaluación operativos que maximicen la transparencia y permitan controlar posibles sesgos vinculados, en muchos casos, con las perspectivas individuales de los investigadores.[10] Más allá de esto, se aboga también por reforzar los mecanismos intersubjetivos de evaluación, es decir, que el acto de evaluar no sea responsabilidad exclusiva de un individuo, sino que haya espacios de discusión grupal que, evidentemente, contribuirían a atenuar posibles arbitrariedades. Pero hay un elemento adicional que tiene que ver con la factibilidad: tanto las evaluaciones cualitativas caso por caso, como con las instancias plenarias de evaluación, dado el volumen que actualmente ha alcanzado el material objeto de evaluación, resultan muchas veces difíciles de implementar. Esto torna complejo, desde el punto de vista operativo, la aplicación de criterios que, desde un punto de vista teórico, metodológico y procedimental, podrían significar una mejora en las prácticas de evaluación. Por otra parte, hay que tener en cuenta los enormes costos económicos que conlleva sostener una estructura de evaluación profesionalizada capaz de afrontar la carga que actualmente ha alcanzado esta tarea y que, todo indica, seguirá incrementándose.

8. Los comentarios precedentes en relación con los límites metodoló-

[10] Por ejemplo, en la convocatoria 2004 para categorización de investigadores en el Programa de Incentivos de la SPU, para asignar las categorías 1 y 2 había incisos que establecían un requisito cualitativo: "haber formado recursos humanos" y "haber contribuido a la formación de recursos humanos", respectivamente. Pero su interpretación generó enormes disparidades entre regiones, disciplinas e incluso comisiones. Muchas veces los evaluadores tenían una idea personal de lo que significaba "formar o contribuir a la formación de recursos humanos" y la aplicaban en la evaluación, aunque no coincidiera en lo más mínimo con la apreciación que tenían otros colegas. En la categorización siguiente, la de 2009, se decidió reemplazar este criterio por otro cuantitativo con el fin de evitar los problemas operativos registrados en la convocatoria anterior.

gicos y fácticos de la evaluación cualitativa de ningún modo implican que aquella cuantitativa esté libre de problemas, más allá de sus supuestas ventajas en términos de replicabilidad, tratamiento uniforme de casos y mayor rapidez de implementación. Mucho se ha discutido en torno de las propiedades conceptuales, las dimensiones y los indicadores que actualmente se utilizan para dar cuenta de las prácticas académicas en el marco de las evaluaciones cuantitativas. Con respecto a algunos indicadores se ha puesto en duda su validez. Y también se ha cuestionado la fiabilidad de los resultados que se obtienen a partir de sus más frecuentes definiciones operativas. Tomemos el siguiente caso como ejemplo: uno de los criterios principales para evaluar la trayectoria académica consiste en considerar la producción científica. Para ello se suelen tener en cuenta dimensiones como su "calidad" e "impacto." Y se recurre a indicadores como: "cantidad de artículos publicados en revistas indizadas de nivel 1, 2, etc.". En esto encontramos varios problemas metodológicos. El primero es que la dimensión que se quiere conocer (por ejemplo, calidad o impacto) se refiere a una unidad de análisis (artículo académico) que no coincide con la unidad de análisis de la cual es atributo el indicador propuesto, porque para juzgar la calidad e impacto del trabajo, como es ampliamente sabido, se consideran atributos de la revista en la que fue publicado. Para defender esta operación se ha argumentado que si un artículo fue publicado en una revista de primer nivel su calidad puede darse por descontada, ya que para aceptarlo se han aplicado estrictos sistemas de referato. Pero si bien podría hacerse esta concesión, aún cuando el conocido caso Sokal pone en duda el argumento, no parece razonable extenderlo especularmente a la situación inversa: es decir, concluir que un artículo que no fue publicado en una revista de ese tipo ostente, por lo tanto, una inexorable "baja calidad" científica o académica. Por otra parte, la cuestión del impacto también resulta problemática. Aun dejando en suspenso la discusión en torno de este criterio, las formas operativas de determinarlo dejan mucho que desear. Por un lado, porque suele tomarse como referente el factor de impacto de la revista, pero no el impacto individual del artículo. Si se consideran en cambio las citas independientes, muchas veces nos encontraremos con la sorpresa de que un artículo publicado local o regionalmente, en una revista Nivel 3, o incluso no indizada, ha tenido mayor "impacto" que otro del mismo autor publicado internacionalmente en una revista Nivel 1. Y esto sin considerar otro tipo de impacto cualitativamente

más sustantivo, como aquél que pueda tener un trabajo relevante a nivel local para explicar una situación social problemática o con potencial transferibilidad de conocimientos para abordarlo eficientemente desde la política pública. Además, si de impacto hablamos, entonces en las ciencias sociales deberíamos reconsiderar el valor que se le asigna a los libros en las evaluaciones, ya que ellos suelen tener mucho mayor volumen de citas que los artículos de revista de los mismos autores. Por supuesto que considerar el nivel de citas también puede resultar engañoso, ya que este indicador, en algunos casos, da cuenta de la calidad de un trabajo y, en otros, implica lo contrario. En efecto, se ha señalado que un libro o artículo puede ser muy citado porque sus pares lo han considerado de alta calidad, y lo toman como referente, o porque lo juzgan un ejemplo de pésima ciencia. Por lo tanto, calidad e impacto no necesariamente están correlacionados. Este problema nos conduce a otra cuestión, que es que no todas las propiedades conceptuales y dimensiones relevantes con relación nuestras actividades y productos académicos pueden ser medidas en sentido estricto. No obstante, en nuestras prácticas de evaluación tendemos a forzar la cuantificación de componentes claramente cualitativos y a reproducir lo que la critica metodológica ha llamado "estiramiento semántico" del concepto de medición, que implica que en nuestro afán por cuantificar, cualquier procedimiento de asignación de una etiqueta numérica implica una medición.

Finalmente, quisiera retomar una afirmación introducida en la segunda sección de este artículo: la evaluación también debe ser entendida como parte constitutiva de la política científica y de educación superior y, por lo tanto, resulta relevante dedicar esfuerzos a su análisis. En línea con los aportes publicados en Camou, Krotsch y Prati (2007), es importante imaginar y emplear dispositivos de evaluación de la evaluación para poder diagnosticar qué efectos y múltiples consecuencias tienen las lógicas de evaluación imperantes en las trayectorias individuales de los académicos, en las instituciones y en el sistema científico y universitario. Este ejercicio, además, no debería limitarse –a mi juicio– a un mero diagnóstico descriptivo, sino que también tendría que influir críticamente sobre las políticas públicas referidas a las áreas en cuestión.

Bibliografía

Burawoy, M. (2005). For Public Sociology. *American Sociological Review*, *70*(1), 4-28.

Bynner, J. & Chisholm, L. (1998). Comparative youth transition research: methods, meanings and research relations. *European Sociological Review,* 14, 131-150.

Camou, A., Krotsch, P. & Prati, M. (2007). *Evaluando la evaluación: políticas universitarias, instituciones y actores en Argentina y América Latina.* Buenos Aires: Prometeo.

Fernández Lamarra, N. & Marquina, M. (Comps.) (2012). *El futuro de la profesión académica. Desafíos para los países emergentes.* Saénz Peña: EDUNTREF.

Fideli, R. (1998). *La comparazione.* Milán: Franco Angeli.

Krippendorff, K. (1980). *Content analysis. An introduction to its methodology.* Newbury Park: Sage.

Hanafi, S. (2011). University systems in the Arab East: Publish globally and perish locally vs. publish locally and perish globally. *Current Sociology, 59*(3), 291-309.

Lazarsfeld, P. (1973). De los conceptos a los índices empíricos. En R. Boudon & P. Lazarsfeld. *Metodología de las ciencias sociales* (Vol. I). Barcelona: Laia.

Marradi, A. (1990). Classification, typology, taxonomy. *Quality & Quantity: International Journal of Methodology*, *24*(2), 129-157.

Marradi, A. (1996). Metodo come arte. *Quaderni di Sociologia*, 10, 71-92.

Piovani, J. I. (2006). *Alle origini della statistica moderna.* Milán: Franco Angeli.

Piovani, J. I. (2008). The historical construction of correlation as a conceptual and operative instrument for empirical research. *Quality & Quantity: International Journal of Methodology, 42*(6), 757-777.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103, 677-680.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Measures in the Social Sciences.* Chicago: Rand McNally.