

Peter Dahler-Larsen*

¿DEBEMOS EVALUARLO TODO? O DE LA ESTIMACIÓN DE LA EVALUABILIDAD A LA CULTURA DE LA EVALUACIÓN**

Vivimos en la era de la evaluación. ¿Debe evaluarse todo en todo momento? En el mundo de la evaluación ha habido un cambio interesante en cómo se responde a esta pregunta. De acuerdo con la antigua idea de la «estimación de la evaluabilidad», una evaluación sólo debe ser llevada a cabo cuando cabe esperar racionalmente que produzca mejoras. Hoy en día, sin embargo, las ideas dominantes de «capacidad de evaluación» y «cultura de la evaluación» sugieren que deben ser promovidos normas y valores que favorezcan la evaluación. Al mismo tiempo, se están generando nuevas categorías de evaluación. El artículo estudia las consecuencias de estos cambios en la forma de responder a la pregunta: ¿se debe evaluar todo?

Palabras clave: políticas públicas, evaluación, regulación.

Clasificación JEL: H43.

1. Introducción

En cada era hay unas cuantas características que habrán de ser recordadas en el futuro. Quizá la nuestra será la «era de la evaluación». Al menos resulta acertado decir que una ola de evaluación sin precedentes está recorriendo muchos países, muchos sectores de actividad pública y muchos niveles de gobierno. Ya en la forma de «cultura de evaluación», capacitación para evaluar o regímenes de evaluación, se observan mu-

chos intentos hoy en día para hacer de la evaluación algo aceptado e institucionalizado. Obviamente esto pone sobre el tapete la cuestión de si es apropiado evaluar todo en todo momento. ¿Existe un algoritmo suficientemente razonable que nos ayude a decidir cuándo evaluar y cuándo no hacerlo? En este artículo trataré esta cuestión a dos niveles. El primer nivel se refiere a la pregunta en sí misma, esto es, desde un punto de vista operativo y práctico. En el segundo nivel, centraré mi atención en cómo el mundo de la evaluación (libros, teorías y Zeitgeist o espíritu de la época) ha respondido esta cuestión a lo largo del tiempo. Mi propuesta es, quizá un tanto provocativa, que este mundo ha pasado desde una concepción modesta del valor de la evaluación a otra autosuficiente.

* PhD. Department of Political Science and Public Management. University of Southern Denmark.

** Traducción de Francisco Luis de Vera Santana.

Primeramente, es necesario, sin embargo, hacer algunas aclaraciones sobre las varias perspectivas que se pueden aplicar cuando hablamos de evaluación.

2. Perspectivas de la evaluación

Algunas discusiones que tienen lugar sobre la evaluación implican confusiones, quizá porque no se clarifican las perspectivas diferentes que están debajo de cada enfoque. Por ejemplo, se puede aplicar una perspectiva prescriptiva o descriptiva. Aquellos interesados en las fórmulas prescriptivas buscan asesoramiento instrumental sobre qué hacer ante una situación dada. Las perspectivas descriptivas no siempre aportan este asesoramiento porque están más orientadas hacia cómo se hace la evaluación realmente, no en cómo debería hacerse. Si entendemos cómo se realiza la evaluación en la práctica, nuestro asesoramiento puede resultar más sensato, aunque no siempre. Es posible que una investigación desapasionada, objetiva e incluso cínica del funcionamiento real de la evaluación no es lo más apropiado para promover la idea de evaluación y para recomendar ciertos de sus instrumentos con entusiasmo. Como investigador interesado en la evaluación, claramente experimento la tensión entre estas dos perspectivas.

Otra diferencia de perspectiva tiene que ver con el hecho de que algunos entienden que interesarse por la evaluación es una elección individual. Otros ven la evaluación como una práctica social, política y organizacional que, aunque hecha por seres humanos, no está definida, formalizada y perfilada por los individuos. Por el contrario, la evaluación es un síntoma de patrones culturales y políticos. Los defensores de esta última perspectiva no pueden hablar de la decisión de evaluar sin pensar en los marcos colectivos e institucionales en los que la evaluación se hace obligatoria.

Siguiente asunto, los investigadores, evaluadores, profesionales y estudiantes de la evaluación mantienen distintos supuestos sobre el grado de racionalidad de los seres humanos. Bajo estrictos supuestos de racionalidad, la evaluación es un instrumento que los seres humanos

diseñan deliberadamente para producir conocimientos con los que ayudar a resolver problemas específicos en situaciones específicas. Sin estos supuestos, la evaluación puede ser una moda pasajera o un ritual institucional, y los resultados de la evaluación inciertos.

Por último, pero no menos importante, la gente tiene supuestos variados sobre el grado de linealidad del proceso de evaluación. Algunos asumen que, una vez que el propósito de una evaluación ha sido decidido, el proceso evaluatorio se desenvolverá de tal manera que el uso real de la evaluación corresponderá con el propósito estipulado. Sin el supuesto de la linealidad del proceso evaluatorio, el propósito estipulado no nos ayuda a predecir su uso real. Por el contrario, el uso de la evaluación puede ser menor, mayor y/o cualitativamente diferente del propósito oficial de la evaluación. Los procesos evaluatorios son abiertos y bastante impredecibles.

Muchos manuales sobre la evaluación al uso adoptan una visión prescriptiva, asumen que los individuos toman decisiones racionales sobre procesos de evaluación predecibles. Es posible, sin embargo, analizar la evaluación como un fenómeno social bajo el supuesto de que la evaluación no es lineal y en gran medida es el resultado de factores irracionales.

Aunque estas perspectivas pueden entenderse entre ellas, no son fácilmente puestas bajo unos mismos estándares. Cuando se clarifican las perspectivas, sin embargo, se hace más fácil entender por qué gente distinta que desarrolla papeles diferentes concibe las evaluaciones de manera tan desigual. Como profesor, y a veces consultor, me descubro a veces tratando de hacer afirmaciones prescriptivas sobre cómo podría ser el buen uso de la evaluación si los individuos fueran racionales y los procesos relativamente predecibles. Como investigador, constantemente veo la necesidad de describir procesos de evaluación complicados en marcos institucionales complejos sin imponer supuestos de racionalidad ni en el nivel micro ni en el nivel macro. La mayor parte de lo que se dice en este artículo, especialmente sobre cómo han cambiado las normas relativas a la evaluación, está basado en esta última perspectiva.

3. La era de la estimación de la evaluabilidad

La estimación de la evaluabilidad es un proceso que lleva a decidir si es sensato evaluar en las circunstancias presentes. La idea es práctica y útil si uno quiere prescripciones sobre cuándo evaluar y cuándo no. Pero la idea no es tan popular hoy en día como lo fue en los setenta (Smith, en Mathison (ed.), 2005: 137). De hecho, el nacimiento y declive de la idea de la estimación de la evaluabilidad es un desarrollo importante en la historia de la evaluación.

La principal cuestión en la estimación de la evaluabilidad no es si se puede hacer evaluación, sino si es racional hacerlo bajo las circunstancias del momento a la luz de las mejoras esperadas de la evaluación (Shadish, Cook y Leviton, 1991: 237). Según la idea de la estimación de la evaluabilidad, un cierto número de circunstancias del evaluando y su contexto debe ser clarificado antes de que se decida hacer la evaluación. Más específicamente, las siguientes cuestiones deben ser respondidas (Shadish, Cook y Leviton, 1991: 237; Rossi, Lipsey y Freeman, 2004: 137):

a) ¿Existe una descripción clara del programa? Si no es así, se hará un mejor uso de los recursos clarificando el programa que evaluando un programa confuso.

b) ¿Está bien implementado el programa? Si se sabe que no es el caso, puede ser más inteligente mejorar su implementación antes de evaluarlo. Las capacidades de obtener conclusiones claras sobre el programa mejoran notablemente si los problemas de implementación han desaparecido y no son responsables del fracaso del programa.

c) ¿Existe una buena teoría del programa? Si hay, es mejor clarificar la lógica del programa e incluso mejorarla antes de evaluar.

d) ¿Existen metas plausibles y bien descritas? Si no existen, el resultado de la evaluación es predecible antes de que se realice. Hay mejores formas de gastar los recursos.

e) ¿Hay datos relevantes accesibles? Si no hay, los recursos de la evaluación podrían gastarse mejor en alternativas a la evaluación.

f) ¿Se han identificado oportunidades para mejorar el programa? Si los usuarios para los que se realiza la evaluación no son capaces o no quieren usar los resultados de la evaluación, la evaluación no va a marcar ninguna diferencia en la práctica.

En esencia, la idea de examinar la evaluabilidad de un programa es altamente racional. Parte del convencimiento de que la evaluación debe ser utilizada solamente en aquellas situaciones en las que pueda ser de utilidad marcando una diferencia. La evaluación debe realizarse solamente cuando sea conveniente, y existen definiciones estrictas de lo que se entiende por conveniente.

Sin embargo, quizá precisamente debido a este supuesto subyacente de estricta racionalidad, esta idea ha ido decayendo desde los años setenta hasta nuestros días (Smith, 2005: 139). Una visión un tanto más cínica es la de que el concepto de estimación de la evaluabilidad se ha desvanecido entre los evaluadores, porque es comercialmente perjudicial renunciar a los honorarios, cosa que sería obligada si el resultado de la estimación resultara negativa (Shadish, Cook y Leviton, 1991: 237). En vez de enfrentarse con un conjunto de criterios estrictos para juzgar si una evaluación puede ser vendida, los consultores adoptan ahora el enfoque de vender no solamente una evaluación sino una cultura de evaluación, sin discutir mucho sobre su legitimidad.

Cualquiera que sea la razón, en mi opinión resulta claro que la estimación de evaluabilidad ha sido en la actualidad reemplazada por ideas, normas y prescripciones que están basadas en el valor de la evaluación en general, más que en las características específicas de la situación a evaluar de la era anterior.

4. Cultura de la evaluación y fortalecimiento de las capacidades evaluatorias

El fortalecimiento de las capacidades evaluatorias se dirige a generar un estado de cosas en el que la evaluación cualitativa de los programas y el uso apropiado de sus resultados son prácticas habituales (Baizerman,

Compton y Stockdill, 2005: 38), siendo así que bajo esta concepción el énfasis se pone en la existencia de los factores que soportan y fortalecen esta situación potencial, no en la realización de evaluaciones.

La idea de cultura de la evaluación y de capacidades evaluatorias es consistente con tendencias de profundo calado en la moderna cultura occidental. En la era reciente, que algunos sociólogos describen como «modernización reflexiva», el progreso automático es una creencia en declive. En vez de eso, la atención que se le ha prestado a los efectos secundarios, la complejidad, y el criticismo ha preparado el terreno para que se demanden mecanismos organizacionales que prometan mejorar el aprendizaje, la retroalimentación y la dirección bajo condiciones complejas.

La evaluación encaja bien en este clima social que apoya los mitos del control, la auditoría y la evaluación (Dahler-Larsen, 2001; Power, 1997). La idea dominante hoy en día es que la evaluación debe ser institucionalizada y aceptada. Debe ser una parte integrante de las organizaciones modernas.

Dentro de los factores que fortalecen la capacidad de evaluación se encuentran aspectos culturales y estructurales de las organizaciones. Desde el lado estructural, la evaluación se basa en la capacidad de una organización para colocar la evaluación en un lugar adecuado en la estructura vertical y horizontal de la organización y en asegurar el apoyo gerencial, etcétera. También incluye la capacidad de contar con personal competente para realizar evaluaciones de calidad dentro de los límites presupuestarios asignados y basadas en las reglas de la organización y en otros mecanismos formales e informales de coordinación. Más del lado cultural, la evaluación se apoya en las normas, valores y hábitos favorables a la evaluación. No existe un consenso absoluto sobre estas normas, pero pueden incluir, por ejemplo, costumbres que obligan a disponer de documentos escritos de los métodos y los objetivos, indicadores existentes de procesos y resultados, la preferencia por experimentar, la tolerancia de los errores, el compartir los conocimientos, los foros de aprendizaje organizacional, etcétera.

Un ideal importante es el crear organizaciones auto-suficientes en materia de evaluación. Esta idea ha ganado terreno recientemente. En algunos contextos se ha convertido en un ideal que se entiende ha de acompañar necesariamente a toda labor de reforma. La actividad de fortalecer las capacidades evaluatorias, curiosamente, es una actividad de duración infinita en la que no se termina nunca de producir reglas, procedimientos, prácticas y políticas (Baizerman, Compton y Stockdill, 2005: 38); esto la hace diferente de la actividad de ejecutar una evaluación, que es por definición una actividad finita. Como consecuencia de haber pasado de la estimación de la evaluabilidad al fortalecimiento de las capacidades de evaluación, el mundo de la evaluación ha dejado de centrar su interés en proyectos que deben ser terminados para enfocarse en proyectos que en principio no tienen un punto final lógico.

Antes de entablar una discusión crítica sobre las características de la cultura de la evaluación y del fortalecimiento de las capacidades evaluatorias, hay que reconocer primero los méritos de la idea central.

Con frecuencia se ha detectado que la barrera más importante para que se saquen enseñanzas de las evaluaciones realizadas no es la calidad técnica de las evaluaciones, sino las normas y valores no favorables a la evaluación que puedan existir entre los que toman las decisiones así como entre los grupos interesados. Como consecuencia, fortalecer las capacidades evaluatorias en el seno de las organizaciones parece ser una estrategia acertada (al dar este paso, los defensores de la cultura y el fortalecimiento de las capacidades sostienen una visión sobre el contexto político y organizacional que rodea al evaluando más avanzada que la que mantienen los defensores de la estimación de la evaluabilidad. Mientras que los últimos consideran el contexto como un factor contingente que les viene dado, los primeros lo ven como un objeto de intervención).

La idea de capacidad evaluatoria y de cultura de evaluación es, al menos parcialmente, también el resultado de observar que los estudios de evaluación realizados por expertos externos no suelen tener impactos de largo

alcance en las organizaciones. Los conceptos de cultura de la organización y fortalecimiento de capacidades se enfocan por tanto en la capacidad o voluntad de una organización para evaluarse más que en la finalización de una evaluación aislada. Puede resultar inteligente invertir en esta capacitación para el futuro por las razones aquí expuestas.

Sin embargo, la idea de cultura de evaluación y capacitación no deja de presentar problemas. Primero que nada, no existe que yo sepa evidencia arrolladora que pruebe que las organizaciones con cultura de evaluación y capacidad de evaluación tengan un desempeño mejor que el resto de las organizaciones en base a unos indicadores objetivos. A continuación, la atención especial que ha recibido el crear un clima favorable a la evaluación, en la práctica puede implicar que las recomendaciones generadas por lo que se conoce como estimación de la evaluabilidad puedan ser desechadas. Bajo un clima de cultura de evaluación, amistoso para la evaluación, las evaluaciones pueden ser realizadas en situaciones que no resultarían recomendables bajo un estricto análisis racional de la situación. Si la evaluación se convierte en la norma general, la contribución promedio de cada evaluación puede verse disminuida en la medida en que se harán evaluaciones a un abanico mayor de situaciones con independencia de la evaluabilidad de un evaluando en una situación dada.

De hecho, no hay razón para pensar que la evaluabilidad general de las intervenciones públicas se haya incrementado automáticamente con el paso del tiempo. En muchas situaciones, lo opuesto puede ser el caso. Si las intervenciones públicas son cada vez más complejas, con impacto en varios niveles de gobierno y que se componen de paquetes de instrumentos, no puede decirse que sean más fáciles de evaluar (al menos sin previamente realizar un esfuerzo extraordinario en clarificar la intervención, la teoría del programa, los objetivos, las necesidades de información, y las opciones para la mejora del programa). De hecho, algunos evaluadores hablan de intervenciones que son tan complejas que no admiten su evaluación.

Adicionalmente, hemos visto en años recientes un aumento de los tipos de intervenciones que pueden ser denominadas «regulaciones de regulación». Con esto me refiero a tipos de intervención dirigidas al control y encauzamiento de otras intervenciones. La regulación de las regulaciones persigue directa o indirectamente bien influir el diseño de otras regulaciones bien motivar a los autores o ejecutores de otras regulaciones para que piensen o se comporten de una cierta manera. La regulación de las regulaciones es a menudo compleja y se extiende sobre áreas de actividad amplias, sobre distintos niveles de decisión política y sobre diversos instrumentos de política.

La respuesta ofrecida hoy en día desde el mundo de la evaluación es recomendar que la cultura de la evaluación y la capacitación deben ser un elemento esencial del complejo sistema que realmente realiza la regulación de la regulación. Ésta es, por supuesto, una respuesta sensata, pero es también un tanto arriesgada, porque contesta a una abstracción por medio de otra abstracción.

Existe un riesgo claro que enfocar la atención sobre capacidades, valores, y potencialidades abstractas se convierta en dominante, mientras que se las mejoras reales propuestas por evaluaciones específicas pierda atención. Esta hipótesis se apoya en la observación que las agencias que controlan organizaciones a veces centran su atención en la «capacidad y cultura evaluadora» de la organización bajo su control que en la realización fáctica de evaluaciones.

El desplazamiento desde la estimación de la evaluación a la cultura y la capacidad de evaluación ha tenido importantes consecuencias en el uso de la evaluación. En la era de la estimación de la evaluabilidad, lo esencial era el uso instrumental de la evaluación. Una estimación de evaluabilidad se daba por terminada cuando era capaz de identificar opciones concretas para mejorar el programa, lo cual era condición necesaria para empezar la evaluación real. Paralelamente, el principal problema con el uso de la evaluación en esa época (al menos tal como se mostraba en el debate sobre la evaluación en los años setenta) era una aparente falta de uso instrumental de las evaluaciones.

El hecho de que la evaluación es a menudo recomendada de manera general en nuestra era incluso bajo circunstancias no evaluables (que no era el caso en la era de la estimación de la evaluabilidad) es una indicación de que el mundo que nos ocupa ha ganado más confianza en sí mismo. Las ideas dominantes sobre la evaluación de nuestra era no insisten ya en que sea necesario razonar lógicamente el valor de cada una de las evaluaciones. Podemos por tanto hacer la hipótesis de que la evaluación será utilizada en ciertas situaciones en donde la probabilidad de que sea usada instrumentalmente en la mejora de la situación sea baja.

En las situaciones en donde la cultura de la evaluación ha sido impuesta, el correspondiente uso de la evaluación puede ser de «uso obligatorio» (Weiss, Murphy-Graham y Birkeland, 2005). Casi con certeza, si la estimación racional de la evaluabilidad hubiese sido el punto de partida, el uso obligatorio seguiría unos patrones diferentes a los que se hubiesen previsto.

Por tanto, con la emergencia de la idea de la cultura y capacidad de evaluación —y la creciente confianza en sí mismo del mundo de la evaluación— el problema fundamental respecto a uso de la evaluación ya no es el de la no utilización. En concreto, el carácter del problema del uso de la evaluación ha cambiado.

5. Nuevas categorías de uso

Recientemente el mundo de la evaluación ha hecho incursiones en formas nuevas y no convencionales del uso de la evaluación. Para aclarar este punto, es necesario primero describir unas cuantas categorías convencionales.

Primeramente, está el «uso instrumental», que para mí significa formas de uso que son idénticas a los propósitos formales de la evaluación (tales como rendición de cuentas o aprendizaje) y que están estrechamente ligadas en tiempo y lugar con la evaluación en sí misma, sugiriendo así un alto grado de linealidad en el proceso de evaluación.

A continuación, hay formas de utilización políticas, estratégicas y tácticas. Aunque no es evidente, entiendo que los procesos de evaluación comprendidos en estas formas de utilización son relativamente lineales y predecibles. Lógicamente, la utilización estratégica y táctica de la evaluación es por definición una expresión de intereses subyacentes, estructuras de poder e intenciones ocultas, que explican de esta manera la utilización política de la evaluación.

Lo que ha aparecido recientemente, sin embargo, es una visión más radical de lo que se llamarían formas no convencionales de la utilización de la evaluación. Se entiende por formas no convencionales de la utilización de la evaluación a los cambios que tienen lugar en la realidad social de manera menos lineal y menos racional que las que sugieren las utilidades convencionales. Las formas no convencionales de utilización tienen lugar porque los procesos de evaluación y/o los resultados/indicadores/impactos tienen vida propia que va mucho más allá de los propósitos formales o informales de la evaluación. Estas formas de utilización es probable que tengan lugar con mayor frecuencia en situaciones en las que la evaluación se pone en práctica por imposición normativa más que porque se esperara de manera racional que fuera usada de manera instrumental.

Se ha descubierto que los procesos evaluatorios agitan ciertos procesos organizacionales y sociopsicológicos, la llamada «utilización de proceso». En una encuesta de aproximadamente 150 directores de servicios sociales de los municipios que describieron la evaluación más reciente realizada en su área de actividad, los encuestados informaron de un cierto número de utilidades de la evaluación que estaban más allá de las consecuencias inmediatas para el evaluando. «Utilización de proceso» en estos casos significa que, por ejemplo:

— La evaluación centra la atención en una intervención en particular y en las personas vinculadas a ella desde gerentes a socios externos.

— La evaluación prepara a la gente para el cambio y deja claro que hay posibilidades de hacer mejoras;

— La evaluación lleva a un intercambio de perspectivas y experiencias entre los socios y los demás interesados.

— La evaluación crea debate.

— La evaluación crea entusiasmo.

— La evaluación crea frustración entre aquellos que no participan en ella en condiciones de igualdad.

Estos hallazgos sugieren que la evaluación puede jugar un papel importante en la distribución de energía y dedicación en las organizaciones. Esto es teóricamente significativo, porque la teoría nos dice que la energía y la dedicación pueden ser recursos escasos en las organizaciones y su asignación es muy importante en la práctica.

No solamente los procesos de evaluación, sino que también los criterios de evaluación y los indicadores pueden tener vida propia. Déjenme ilustrar este punto refiriéndome a lo que Thiel y Leew (2002) llaman «fijación con la medición». Fijación con la medición ocurre cuando los profesionales centran su atención en qué se está midiendo como un indicador de calidad, frecuentemente sin preocuparse de la auténtica calidad. Esto puede ser denominado como una forma «trivial» de fijación con la medición porque lo que se persigue es la auténtica calidad, objetivo del que desvía la fijación con la medición. En este sentido, la fijación con la medición es no intencionada. Opuestamente a esto, en el caso de fijación con la medición avanzada, el indicador ofrece una definición de calidad junto con un indicador de cómo medir la calidad. Por ejemplo, «lo que medimos con un *test* de inteligencia es la inteligencia». O bien, «nadie sabe lo que es la calidad escolar, pero en nuestro país pensamos que los buenos resultados en los exámenes constituyen un aspecto importante de la calidad escolar».

Con fijación con la medición avanzada no es posible demostrar una distinción entre calidad auténtica y calidad medida por un indicador, toda vez que el segundo ayuda a definir la primera. Este es un efecto constitutivo que debe ser tomado seriamente y que no se entiende apropiadamente cuando se le caracteriza como «patoló-

gico» o «no intencionado». Lo peculiar es que mientras el indicador se construye socialmente, el fenómeno que supuestamente mide también está construido socialmente. Por lo tanto, es insuficiente criticar la medición porque supone considerar el fenómeno como invariable o conocido por todos. Por el contrario, los criterios de evaluación ayudan a definir el marco en el que las actividades de evaluación adquieren sentido.

Los efectos constitutivos deben ser estudiados no solamente por lo que no son (no intencionados, sin propósito, insalubre, irracionales), sino que también merecen ser estudiados por lo que son.

6. Efectos constitutivos de la evaluación

Los criterios de evaluación definen los marcos interpretativos y las visiones del mundo

De manera muy fundamental, los criterios de evaluación son una especie de «herramienta para dar sentido a las cosas» (Mark, Henry y Julnes, 2000), ofrecen claves interpretativas que captan la atención, definen discursos y orientan las acciones en direcciones determinadas. Su función más básica no es verificar qué pasa en un área particular de actividad, sino construir una definición de esa actividad de tal manera que se pueda verificar.

Un ejemplo ilustrativo es una encuesta que le permite a los directivos comprobar cómo sus familias evalúan su presencia en la familia (Hochschild, 2004). Aunque presentada como una herramienta progresiva para recordar la importancia de la vida familiar a los directivos, la encuesta también contribuye a identificar la vida familiar como algo reducido a un número limitado de dimensiones racionalizadas, cuantificables y gestionables. También introduce la idea de que la persona sujeta a la encuesta puede o debe cambiar su comportamiento familiar con el objetivo de influenciar los resultados de su encuesta.

El conocimiento evaluatorio puede así colonizar áreas de la vida social especificando y cuantificando sus valo-

res inherentes de manera explícita cosa que, en su defecto, no sería el caso. Un efecto constitutivo y transformativo (y por tanto amenazador) similar del conocimiento evaluativo sobre cómo las comunidades se ven y explican a sí mismas es analizado por Schwandt y Dahler-Larsen (de próxima publicación).

Un aspecto constitutivo de los criterios de evaluación se refiere no a cómo se usan en la medición y verificación, sino a cómo ayudan a definir el contenido de lo que está siendo evaluado.

Los criterios de evaluación definen «contenido»

Los criterios de evaluación definen lo que es central en el trabajo. Los criterios de evaluación ayudan a determinar qué actores deben esforzarse para llevar a cabo ciertas actividades. Otras veces, por el contrario, la evaluación provoca la formulación de criterios de éxito explícitos a través de prácticas con formas intuitivas e implícitas (Munro, 2004: 1082). Puesto que en muchas áreas de actividad la primera vez que se hace esto es con ocasión de la evaluación, la evaluación tiene un efecto constitutivo en la definición de las dimensiones centrales de las actividades que componen el trabajo.

Cuando la labor de examinar se convierte en una práctica social extendida, puede suceder que ponga en marcha un montón de maneras llamativas de reaccionar entre los profesores y alumnos por su capacidad para definir qué cuenta en lo que se entiende por una enseñanza exitosa. Un concepto como el de «enseñar para aprobar» describe alguna de estas reacciones. Como ejemplo, realizar un ensayo de examen en el que el profesor incluye muchas de las preguntas que sabe que figurarán en el examen real. Después del ensayo los alumnos tienen la posibilidad de ver las respuestas correctas. En el examen real la habilidad del estudiante para recordar las respuestas a las cuestiones del ensayo de examen se valora positivamente. La enseñanza consiste en aprender a hacer los exámenes.

Los criterios explícitos a veces son una invitación a «tomar atajos» y a apostar para puntuar mejor en los in-

dicadores establecidos (Courtney *et al.*, 2004). De hecho, cuanto más importancia política e institucional tenga un indicador del aspecto central de un trabajo, más será el indicador objeto de complejos procesos sociales que en ocasiones amenazarán la validez y confianza del indicador, haciendo más difícil entender qué es lo que el indicador significa realmente (Vaulliamy y Webb, 2001). Cuanto más se haya invertido política y organizativamente en el indicador, más difícil será analizar libremente su interpretación.

Cuando son importantes, a veces los indicadores evocan un abanico de actividades irracionales dirigidas a afectar los valores que toman los indicadores (Fountain, 2001: 66), pero no dirigidas a la calidad del desempeño en sí mismo. Sin embargo, sólo si se sabe lo que se entiende por calidad, se puede criticar un indicador particular por su poca validez o práctica «corrupta». En un apartado anterior se hizo referencia a esto como fijación de medición trivial. Por otra parte, con fijación de medición avanzada, los valores de los exámenes se convierten en una forma privilegiada de conocimiento y ayudan a establecer definiciones de calidad socialmente dominantes. Bajo esas circunstancias, los ensayos de exámenes que utilizan preguntas de los exámenes y enseñan cómo aprobarlos podrían ser una manera estratégicamente significativa de mejorar la calidad escolar tal como es definida.

En un caso interesante los políticos implantaron unos criterios de calidad nuevos en una sala de urgencias. Se medían los minutos desde que el paciente entraba hasta que era atendido por alguien del personal. La sala de urgencias reaccionó contratando «enfermeros saludadores» que saludaban a los pacientes a los pocos minutos de que llegaran. No pasaba nada más. Si la reducción de las listas de espera se define como en este ejemplo, los participantes en el sistema de evaluación no eran corruptos sino que se sometieron concienzudamente a este criterio de evaluación.

De hecho, los criterios de evaluación se pueden convertir en metas en sí mismos. Pueden centrar la atención en sus iconos o símbolos de calidad y definir lo que

no se mide como no existente, lo que se denomina visión túnel (van Thiel y Leeuw, 2002). «Calidad» por definición es quizá un concepto relativamente abierto, pero cuando un sistema de indicadores representa la calidad de una cierta manera, la representación pasa a ocupar el lugar de una definición socialmente válida.

Conforme a esto, los criterios de evaluación son plantillas de conocimientos que comunican lo que es deseable políticamente en un área de trabajo que, socialmente, es altamente productiva. Quizás, los políticos, gestores, consultores y administradores no comprenden todas las implicaciones de sus, a veces, poco cuidadas definiciones de algunos indicadores de desempeño. Obtienen lo que quieren sin entender las ramificaciones.

Por ejemplo, en Dinamarca actualmente muchas escuelas centran su atención en las llamadas «calificaciones blandas» y «competencia social» junto a medidas de calidad fácil y convencionalmente cuantificadas, pero estos nuevos conceptos son difíciles de retener. A medida que emerjan estrategias y técnicas para hacer operativo y medir el concepto de «calificaciones blandas», aparecerán definiciones prácticas de este fenómeno y comenzarán a circular por el mundo. Concebir una manera metodológica/técnica de hacer operativo un concepto, en otras palabras, la elección de un indicador, es una acción esencialmente interpretativa con implicaciones socialmente productivas. No se puede criticar por pobre validación en la medida que ello codefine la definición conceptual de qué es lo intenta medir al mismo tiempo que lo mide.

Los criterios de evaluación son constitutivos de identidades y relaciones sociales

Los criterios de evaluación estipulan identidades sociales y sus relaciones mutuas. Para cada categoría social definida por los sistemas de indicadores, sus normas son a menudo establecidas de manera estándar. Un efecto constitutivo simple pero poderoso emana de la comparación de indicadores entre unidades sociales tales como instituciones, grupos e individuos. Mientras

que los buenos estándares son inherentemente difíciles de establecer en base a argumentos lógicos (Stake, 2004), algunos sistemas de indicadores simplemente toman la media aritmética como la norma con la cual deben ser comparadas las unidades individuales. Aunque simple matemáticamente y no sofisticado filosóficamente, este enfoque tiene repercusiones sociales remarcables. Esta manera de hacer comparaciones implica no sólo que las identidades de las diferentes unidades son de la misma naturaleza, sino también que se encuentran en una relación de competición que lleva a cada una de ellas a estar interesada en rebajar las puntuaciones de los competidores porque la media de todos define el estándar.

En una evaluación de desempeño del personal, una distribución forzada de las calificaciones tuvo unos efectos adversos sobre la motivación de la mitad del personal que fue calificado por debajo de la media (McBriarty, 1988). Hay pues relaciones constitutivas interesantes entre las propiedades de las escalas de medición y las relaciones e identidades sociales.

Cuando se hacen comparaciones cuantitativas, las formas en que aparecen diferencias en calificaciones dependen de la naturaleza de la medición. Desde el matemático Mandelbrot se sabe que con sistemas de medición cada vez más sensibles, se pueden hallar diferencias más pequeñas (van Thiel y Leeuw, 2002). Esas diferencias con frecuencia obligan a prestar atención política, gestora y profesional. De esta manera, los sistemas de indicadores de desempeño colaboran a provocar el problema o el riesgo que supuestamente deben monitorizar (Rothstein, Huber y Gaskell, 2006), y reconfiguran las relaciones entre los actores sociales implicados.

Los indicadores y otros criterios de evaluación colaboran en convertir a los actores implicados en categorías tales como «pacientes», «clientes», o consumidores. Los indicadores de la «satisfacción del consumidor» pueden tácitamente promover la identidad de «consumidor» de educación superior en detrimento de otras identidades tales como la de «estudiante» (Cheney *et al.*, 1997). La aparente libertad del consumidor ha

de comprarse pagando un precio. El consumidor no tiene ninguna responsabilidad en el diseño de las políticas. El consumidor es también el objeto de un abanico de mecanismos de control —socialización— y normalización (Bauman, 1983) y sus preferencias se mantienen bajo estricto control cuantitativo mediante los sistemas de indicadores de desempeño.

Quizás de manera paralela, la evaluación también influye la propia definición de un profesional, e incluso reduce el papel del profesional. Frecuentemente, los indicadores estandarizados reducen la libertad profesional en la práctica de la enseñanza. El equilibrio entre profesionales y gestores puede alterarse en función de cómo los sistemas de indicadores enfatizan las maneras en que las organizaciones se controlan a sí mismas y no tanto porque se haya realizado un examen del equilibrio en sí mismo (Munro, 2004: 1079). De hecho, en muchas ocasiones los indicadores basados en la satisfacción del consumidor no se pueden entender en base a una comprensión profunda de las características de las actividades, sino que por el contrario se entienden mejor como una alianza estratégica entre gestores y consumidores en perjuicio de los profesionales.

Las relaciones corporativas entre profesionales tales como profesores son afectadas en la medida en que el sistema de indicadores agrupa a algunos profesores (en escuelas o grupos) y provoca el enfrentamiento de unos grupos contra otros. Los efectos de los indicadores de desempeño sobre la autodefinición, el compromiso y la moral son quizás uno de los asuntos más críticos.

Los criterios de evaluación son constitutivos de marcos temporales

El efecto constitutivo de los criterios de evaluación no solamente se manifiesta en la definición del contenido del trabajo sino que también establece marcos temporales en los que los resultados de una actividad se espera que tenga lugar. Los indicadores con marcos temporales cortos pueden privilegiar algunas actividades a expensas de otras.

Hojlund y la Cour (2001) demostraron cómo la medición digital del tiempo de cuidados de los mayores cambia fundamentalmente tanto el contenido de los servicios como las relaciones entre el cuidador y el receptor. En el mismo orden de cosas, ya hemos mencionado cómo el marco temporal que define qué se espera que un «niño normal» sea capaz de hacer a una cierta edad provoca simultáneamente una distinción entre niños «normales» y «por debajo de lo normal».

Cuando la satisfacción del consumidor se introduce como criterio de evaluación en la educación superior, con frecuencia aparece asociada la idea de gratificación instantánea, y ambos factores tienden a redefinir el papel del profesor respecto a los estudiantes.

El tiempo es frecuentemente una dimensión privilegiada para hacer mediciones, porque es general y abstracta (como el dinero). Por esta razón, los sistemas de indicadores con frecuencia transforman más substantivas dimensiones de la calidad en medidas de rapidez del servicio (Munro, 2004: 1080), pero este cambio tiene efectos en el contenido del trabajo y en las relaciones sociales.

Los efectos constitutivos pueden desplazarse en el tiempo y a través de niveles de análisis

La cultura de exámenes dominante en las instituciones educativas puede tener efectos a gran escala en la cultura y mentalidad de una sociedad. Un interesante estudio sobre el sistema de exámenes en China ha demostrado efectos con un siglo de duración sobre la cultura y la estructura social. Los criterios de exámenes atraparon a todo el país dentro de una estrategia educativa específica que con los años se ha demostrado competitivamente improductiva (Suen, 2003). Los sistemas de exámenes y los regímenes de indicadores provocan los llamados efectos *washback* o influencia del sistema de exámenes en la enseñanza que se imparte (Bayle, 1996) y que se pueden observar en los currículos, los planes de enseñanza y en las prácticas pedagógicas. Entre estos efectos uno podría señalar nuevas solucio-

nes políticas e institucionales. Por ejemplo, en Dinamarca una ola internacional de exámenes educativos facilitó el camino para el establecimiento de un centro nacional de evaluación. Con el paso de los años, este centro expandió sus actividades y se convirtió en un actor significativo en el campo de la educación como tal. Algunos años después, se introdujo el mandato legal de publicar información escolar, y un centro de investigación comenzó a diseñar maneras de medir la calidad de los profesores individualmente. En otras palabras, la interacción entre evaluaciones, efectos constitutivos, nuevas soluciones institucionales y nuevas evaluaciones puede ser muy compleja.

7. Conclusión

Para bien o para mal, la idea de estimación de la evaluabilidad no tiene fuerza en nuestros días. En vez de en ella, creemos en la construcción de capacidad de evaluación y en el establecimiento de culturas de evaluación. Hay buenas razones para este cambio, entre ellas se encuentra el hecho de que la evaluación no mejora las cosas en la práctica si a la larga no se apoya en normas y valores que sean favorables a la evaluación.

Por otra parte, hoy en día las normas, valores, sistemas y estructuras que fomentan la evaluación están próximas a convertirse en metas en sí mismas. He sugerido en este artículo que este cambio está relacionado con un aumento de seguridad en sí mismo que tiene el mundo de la evaluación —una seguridad que no está sustentada por una base de evidencias similar a la que la evaluación a menudo exige de sus evaluandos. Esta seguridad en sí misma de la evaluación de hoy está relacionada con la creencia de que la evaluación es una cosa buena que debe ser apoyada con normas, valores y prácticas —siempre y cualquiera que sean los factores presentes en una situación específica.

He argumentado además que junto con este cambio en la fundamentación normativa de la evaluación, han aparecido nuevas categorías para describir nuevos pa-

trones de uso de la evaluación, tales como uso proceso y uso constitutivo.

Existen los efectos constitutivos de los criterios de evaluación. Son complejos, relacionales y contextuales. Emergen mediante definiciones, distinciones y mediciones que mejoran algunas construcciones sociales de la realidad. Los indicadores no solamente comunican sobre una actividad, sino que metacomunican sobre las identidades y roles de los seres humanos y sus relaciones. Los criterios de evaluación producen etiquetas socialmente relevantes que se «adhieren» a prácticas y personas, y que colaboran en la organización de las interacciones sociales de una determinada manera. La evaluación «hace sociedad» (Woolgar, 2004: 454).

Los efectos constitutivos son un concepto abierto y no determinista. Sugieren que la finalidad de la evaluación no está determinada por los datos ni por las intenciones que hay detrás de las evaluaciones.

Esta observación lleva a la pregunta de «cuándo evaluar y cuándo no evaluar» a otra escala conceptual distinta a la que era relevante en la era de la estimación de la evaluabilidad. En aquella época, el centro de atención era el uso instrumental de las evaluaciones: si su uso instrumental era posible dadas las circunstancias, y qué debía hacerse para promoverlo.

Hoy, en una era que considera la cultura de la evaluación como un ideal, nuestras preocupaciones sobre cuándo evaluar y cuándo no, deben ser de una naturaleza más amplia.

Hoy, una decisión ponderada sobre si una evaluación debe ser iniciada en un contexto dado debería tomar en consideración los efectos no específicos de «uso proceso» tales como saber cómo se distribuye la energía, dedicación, entusiasmo y satisfacción. Asimismo debe incluir consideraciones meditadas sobre cómo los efectos constitutivos de los criterios de evaluación colaboran a redefinir no sólo lo que se mide, sino el propio significado de las intervenciones públicas, y cómo encajan en las culturas, instituciones y relaciones entre los individuos.

Éstas son sin duda algunas consideraciones amplias. Es un importante punto de partida, sin embargo, reconocer que necesitamos una visión más amplia de la evaluación, del uso de la evaluación, y de la pregunta de si evaluamos o no, que la ofrecida por la moderna evaluación.

Referencias bibliográficas

- [1] ANDERSEN, V. N. y DAHLER-LARSEN, P. (2006): «The Framing of Public Evaluation Data. Transparency and Openness in Danish Schools», en BOYLE, R.; BREUL, J. y DAHLER-LARSEN, P. (eds.): *Open to the Public. Evaluation in the Public Arena*, Transaction, New Brunswick.
- [2] BAILEY, K. M. (1996): *Working for Washback: A Review of the Washback Concept in Language Testing. Language Testing*, 13, 3, 257-279.
- [3] BAIZERMAN, M.; COMPTON, D. W. y STOCKDILL, S. H. (2005): «Article on Capacity Building», en MATHISON, S.: *Encyclopedia of Evaluation*, Sage, Thousand Oaks.
- [4] BOYLE, R.; BREUL, J. y DAHLER-LARSEN, P. (eds.) (2006): *Open to the Public. Evaluation in the Public Arena*, Transaction, New Brunswick.
- [5] CHENEY, G.; MCMILLAN, J. J. y SCHWARTZMAN, R. (1997): «Should we Buy the Student as Consumer Metaphor?», *Working paper at: <http://mtprof.msun.edu/Fall1997/Cheney.html>*.
- [6] COURTNEY, M. E.; NEEDELL, B. y WULCZYN, F. (2004): «Unintended Consequences of the Push for Accountability: The Case of National Child Welfare Performance Standards», *Children and Young Services Review*, 26, 1141-1154.
- [7] DAHLER-LARSEN, P. (2001): *Den Rituelle Reflektion*, Syddansk Universitetsforlag.
- [8] DAHLER-LARSEN, P. (2002): *Evaluering Kortlagt*, Systeme, Århus.
- [9] FOUNTAIN, J. E. (2001): «Paradoxes of Public Sector Customer Service», *Governance*, 14, 1, 55-73.
- [10] HOCHSHILD, A. (2004): «Gennem sprækker i tidsfælden», en JACOBSEN, M. H. y TONBOE, J. (eds.): *Arbejdssamfundet*, Hans Reitzel, 109-130, Copenhagen.
- [11] HØJLUND, H. y LA COUR, A. (2001): «Standardiseret omsorg og lovbestemt fleksibilitet - organisationsændringer på et kærneområde i velfærdsstaten», *Nordiske Organisasjonsstudier*, 2, 3, 91-117.
- [12] MARK, M.; HENRY, G. y JULNES, G. (2000): *Evaluation. An Integrated Framework for Understanding, Guiding, and Improving Public and Nonprofit Policies and Programs*, Jossey-Bass, San Francisco.
- [13] MCBRIARTY, M. A. (1998): «Performance Appraisal: Some Unintended Consequences», *Public Personnel Management*, 17, 4, 421-433.
- [14] POWER, M. (1997): *The Audit Society. Rituals of Verification*, Oxford University Press, Oxford.
- [15] ROTHSTEIN, H.; HUBER, M. y GASKELL, G. (2006): «A Theory of Risk Colonization: The Spiralling Regulatory Logics of Societal and Institutional Risk», *Economy and Society*, 35, 1, 91-112.
- [16] SAHLIN-ANDERSSON, K. (2004): *Presentation at the Conference of the Swedish Evaluation Society (SVUF) in Stockholm in May*.
- [17] SCHWANDT, T. y DAHLER-LARSEN, P. (en prensa): *Evaluation and Community*, aceptado para su publicación.
- [18] SCOTT, W. R. (1987): «The Adolescence of Institutional Theory», *Administrative Science Quarterly*, 32, 493-511.
- [19] SMITH, M. F. (2005): «Article on Evaluability assessment», en MATHISON, S. (ed.): *Encyclopedia of Evaluation*, Sage Publications, Thousand Oaks.
- [20] SUEN, H. K. (2003): «Some Very Long-Term and Some Chronic Effects: The Invisible Dimension of the Consequential Basis of Validity», *Paper presented at the 2003 International Symposium on Developing High-stakes Testing*, Seúl, Corea.
- [21] VAN THIEL, S. y LEEUW, F. (2002): «The Performance Paradox in the Public Sector», *Public Performance and Management Review*, 25, 3, marzo, 267-281.
- [22] WEISS, C. H.; MURPHY-GRAHAM, E. y BIRKELAND, S. (2005): «An Alternative Route to Policy Influence», en *American Journal of Evaluation*, 26, 1, páginas 12-30.