Manuel Souto Otero _Universidad de Oxford

Administración&Cidadanía. _1/2006

Evaluación de políticas: una revisión crítica de las definiciones, diseños y métodos

***La bibliografía sobre** la evaluación de políticas es extremadamente voluminosa y desordenada. Este artículo pretende proporcionar un resumen crítico de las definiciones, diseños y métodos utilizados en la evaluación de políticas para aquellos lectores que estén interesados en el tema pero que se sientan desbordados y, lógicamente, confusos debido a la variedad de enfoques existente. El artículo proporciona también una lista de los elementos que caracterizan a la buena evaluación de políticas, que pretende ser de utilidad para evaluadores, administradores de políticas y otras personas implicadas en el uso y examen de las evaluaciones de políticas.&*

& Palabras clave: evaluación de políticas públicas, elaboración de políticas públicas, impacto de políticas públicas, métodos de investigación

1. ¿QUÉ ES LA EVALUACIÓN DE POLÍTICAS?

El análisis de las políticas públicas pretende determinar cuál de las varias políticas públicas o gubernamentales posibles alcanzará unas metas determinadas en mayor medida, en vista de las relaciones entre las políticas y las metas (Nagel, 1990). En esta tarea iterativa, la evaluación de políticas desempeña un papel determinante para ayudar a los gestores a diseñar o mejorar el diseño de políticas, programas e iniciativas, y para proporcionar, donde sea necesario, exámenes periódicos de la eficacia de políticas o programas, de sus impactos y de las maneras alternativas de alcanzar los resultados esperados.

Por tanto la evaluación, al igual que las auditorías internas, la gestión de riesgos y otras herramientas, ayuda a los gestores a trabajar eficazmente en entornos políticos sumamente complejos. Pero la evaluación también puede ayudar a los gestores en el control del rendimiento e informar sobre el mismo, y ayudar a los responsables de la toma de decisiones a examinar los resultados de los programas o las políticas. Esto es lo que distingue la evaluación de la auditoría interna, una función que proporciona garantías sobre la estrategia de gestión de riesgos de un departamento o agencia, un marco para el control de la gestión e información, financiera y no financiera, utilizada para la toma de decisiones y la elaboración de informes (Treasury Board of Canada Secretariat, 2001).

La evaluación de políticas puede definirse como un conjunto de métodos de investigación que pretende "investigar sistemáticamente la eficacia de las intervenciones sociales (...) de forma que mejoren las condiciones sociales" (Rossi *et al.,* 1999: 20). La importancia de la noción de conjunto de métodos de investigación es fundamental en esta definición: la evaluación de políticas utiliza una amplia gama de diseños y métodos, y no privilegia ninguno de ellos; antes bien, reconoce el potencial complementario de diversos métodos de investigación. Los métodos utilizados en la evaluación y análisis de las políticas derivan generalmente de los temas sustantivos que se traten, más que por preferencias *a priori*[1] (Greene *et al.,* 2001).

Este artículo proporciona una revisión de los diferentes tipos y métodos de evaluación. No obstante, el lector debería recordar que, aunque amplia, nuestra revisión no pretende ser exhaustiva, sino que busca ofrecer un resumen y análisis de los principales enfoques de evaluación. Algunos de los tipos y métodos de evaluación presentados aquí son complementarios y otros alternativos; unos pueden aplicarse de manera general mientras otros tienen aplicaciones bastante específicas. El artículo se estructura como sigue: la sección dos presenta diferentes tipos de evaluaciones y de diseños de evaluación y prestando particular atención a las evaluaciones conclusivas, la forma de evaluación más extendida, mientras que la sección tres se centra en los métodos de recolección de datos que pueden utilizarse para llevar a cabo estos diferentes tipos de evaluación; por su parte, la cuarta sección concluye ofreciendo algunos criterios a satisfacer en una buena evaluación de políticas.

2. TIPOS DE EVALUACIÓN

2.1. Introducción

En algunos países europeos hay poca tradición de evaluación de políticas: éstas son diseñadas, puestas en práctica y cambiadas sin estudiar qué resultado podrían haber dado y cuál dieron realmente. En otros países, la tradición de la evaluación está mucho más desarrollada, hasta el punto de que las políticas son evaluadas casi por defecto, independientemente de las

características específicas de las diversas políticas, programas o proyectos. Ninguna de estas situaciones es la ideal.

La primera pregunta que hay que plantear a la hora de decidir la evaluación de una política, un programa o un proyecto es la siguiente: ¿puede siquiera evaluarse esta intervención? Algunas políticas y algunos programas son tan complejos y difusos que hay pocas posibilidades de que satisfagan los tres requisitos principales que posibilitan la evaluación, a saber: que las intervenciones y la población a que están destinadas estén claras y sean identificables; que los resultados sean claros, específicos y conmensurables; y que se pueda poner en práctica un plan de evaluación apropiado (Patton, 2002).

Si se satisfacen estos criterios, la evaluación de políticas debería producir beneficios significativos. Ciertamente, los gobiernos, las organizaciones internacionales y demás partes implicadas hacen uso de una gama de diversos tipos de evaluación, dependiendo de sus objetivos, su enfoque analítico favorito y el estadio de desarrollo concreto de la política acerca de la cual se requiere información.

En adelante se presentan varios tipos de evaluación, que incluyen métodos de valoración rápida, análisis lógico del contexto, indicadores de rendimiento, evaluaciones antes, durante y después de la puesta en práctica, evaluaciones basadas en las metas e independientes de ellas, evaluaciones basadas en la teoría, evaluaciones cuantitativas y cualitativas, evaluaciones formativas y conclusivas, informes de seguimiento del gasto público, y análisis coste - beneficio, coste - efectividad y coste - utilidad. Como ya se ha hecho notar, algunos de estos tipos de evaluación son complementarios y otros son exclusivos. Por tanto, la división entre tipos que aquí se presenta no pretende ser una clasificación de tipos de evaluación: por ejemplo, no hay un *fundamentum divisionis* compartido (Marradi, 1990) por todos los tipos de evaluación aquí presentados.

2.2. Métodos de valoración rápida, marco lógico e indicadores de rendimiento

Los modelos presentados en esta subsección no son métodos exhaustivos de evaluación, pero tienen en común con las evaluaciones su intención de contribuir a los procesos de toma de decisiones.

Los primeros métodos, los métodos de valoración rápida, son formas, rápidas y de bajo coste de reunir los puntos de vista y los comentarios de los beneficiarios de una política, así como de otras partes interesadas relevantes, para dar respuesta a la necesidad de información de los responsables de toma de decisiones. Proporcionan información, normalmente cualitativa, recopilada rápidamente para la toma de decisiones de los gestores, especialmente en el nivel de los proyectos y programas aunque, normalmente, sus conclusiones no pueden generalizarse y son menos válidas, fidedignas y creíbles que las que se obtienen por medio de las evaluaciones. Los métodos de valoración rápida pueden incluir entrevistas con informadores clave (una serie de preguntas amplias formuladas a individuos seleccionados por su conocimiento de un tema y su experiencia en el mismo), discusiones de grupo (un debate entre, por lo general, de ocho a doce participantes con experiencias similares, como por ejemplo los beneficiarios o el personal del programa), entrevistas en grupos de comunidad (una serie de preguntas y un debate que tienen lugar en un encuentro abierto a todos los miembros de la comunidad), observación directa (sistema basado en el uso de un impreso de observación detallado para registrar lo que se ve y oye en el escenario de un programa) o mini-informes (que usan un cuestionario estructurado con un número limitado de preguntas concretas que se le presenta a un número pequeño de personas escogidas al azar o siguiendo algún criterio).

El marco lógico va un paso más allá de los métodos de valoración rápida. Pretende ayudar a aclarar los objetivos de cualquier proyecto, programa o política e identificar las relaciones causa-efecto que cabe esperar –la 'lógica del programa'– en la siguiente concatenación de resultados: recursos, procesos, resultados, consecuencias e impacto (véase la sección 2.6 para una definición de estos elementos). El análisis lógico del contexto es, pues, una manera de implicar a los interesados en la concreción de los objetivos y el diseño de actividades, y conduce a la identificación de indicadores de rendimiento en cada fase de esta concatenación, además de a la identificación de los riesgos que podrían impedir el cumplimiento de estos objetivos.

A su vez, los indicadores de rendimiento son mediciones de recursos, proceso, resultados, consecuencias e impacto de los proyectos, programas o estrategias de desarrollo. Cuando los respalda una recopilación sólida de datos y un informe sobre éstos, los indicadores permiten a los gestores controlar los progresos realizados, demostrar los resultados alcanzados y adoptar acciones correctivas para mejorar la prestación de servicios. El uso de indicadores de rendimiento implica el establecimiento de metas y la valoración de los progresos que se hacen para alcanzarlas. Tienen la ventaja de que, por medio de su uso, si han sido seleccionados con rigor, es posible identificar problemas mediante un sistema de alarmas tempranas que permite que se den pasos para corregirlos. También pueden indicar si una determinada intervención necesita una evaluación o un informe exhaustivos desde el principio.

2.3. Evaluaciones ex-ante, intermedias y ex-post

No hace falta decir que no todas las evaluaciones se efectúan al final de la puesta en práctica de una política, programa o proyecto: el proceso de evaluación debería contribuir a mejoras en la política, así como en del diseño de programas y su puesta en práctica. Rossi, por ejemplo, sugiere que se pongan en práctica métodos diseñados a medida de la fase en que se encuentra el programa, es decir, que se 'adecuen las evaluaciones al programa' (Rossi *et al.*, 1999). Por tanto, además de las evaluaciones ex-post, existen las ex-ante y las simultáneas.

Las evaluaciones ex-ante tienen lugar antes del comienzo de una política, programa o proyecto y contribuyen a establecer claramente la lógica que se seguiría para intentar resolver un problema y los métodos para hacerlo, así como los efectos positivos y negativos que se espera suscite la intervención. Normalmente se tienen en cuenta la relación y la coherencia entre objetivos globales, objetivos específicos y la complementariedad de las medidas que se incluirán en el programa; también se dedica una parte importante del trabajo a desarrollar un sistema apropiado de indicadores que se utilizarán en evaluaciones subsiguientes (Comisión Europea, 1999). Durante las evaluaciones ex-ante puede tenerse en cuenta toda una gama de posibles intervenciones; en consecuencia, las evaluaciones previas son una buena oportunidad para mejorar la planificación de políticas y pueden apoyar con éxito discusiones al respecto desde varios puntos de vista.

Las evaluaciones intermedias se realizan durante el desarrollo de una determinada intervención, y normalmente se centran en el transcurso de la misma hasta un punto dado, destacando particularmente los estadios de planificación y puesta en práctica de la iniciativa en cuestión y prestando menos atención a los resultados o consecuencias de la misma, que pueden ser difíciles de identificar en una fase temprana de la puesta en práctica. Examinan el grado de efectividad alcanzado basándose en los indicadores recopilados durante el proceso de seguimiento de la iniciativa y valoran la calidad y relevancia de éstos. Las evaluaciones ex-ante e intermedias no suelen ser un fin en sí mismas sino una forma de mejorar la calidad y la relevancia de un programa: proporcionan una oportunidad para identificar las reorientaciones del

programa mismo que pueden ser necesarias para garantizar que se consiga los objetivos originales (Comisión Europea, 2000).

Pese a todo, las evaluaciones *ex-post* son el tipo más frecuente de evaluación: examinan si se han alcanzado los resultados esperados de una intervención y proporcionan información necesaria para la planificación o puesta en práctica de programas nuevos o revisados. Las evaluaciones *ex-post* suelen usar datos definitivos de seguimiento y comparar los objetivos esperados con los que realmente se han conseguido, incluyendo los impactos obtenidos (Comisión Europea, 1999). Aun así, no todas las evaluaciones posteriores se centran únicamente en el análisis de los resultados alcanzados en relación con las metas fijadas para la intervención. Así pues, podemos distinguir entre evaluaciones basadas en las metas, evaluaciones independientes de las metas y evaluaciones basadas en la teoría.

2.4. Evaluaciones de metas y evaluaciones libres

Una de las preguntas más frecuentes en la evaluación de políticas es si se han cumplido o no las metas de una política, un programa o un proyecto: el examen que responde a esta pregunta se conoce como evaluación de metas (Patton, 2002). En la bibliografía estadounidense sobre las evaluaciones se conoce también como 'seguimiento legislativo', porque analiza si se dan alcanzado o no los resultados esperados de una política gubernamental.

Sin embargo, los creadores de políticas y los evaluadores están a menudo interesados también en los resultados o efectos imprevistos, positivos o negativos, de una política, programa o proyecto, aun sin saber necesariamente cuáles eran las metas prefijadas. Este tipo de evaluación de políticas es a menudo importante para establecer la relación entre coste y beneficio o coste y utilidad de una política, un programa o una intervención. Scriven (1972), por ejemplo, propugna la 'evaluación libre', porque el evaluador asume la responsabilidad de decidir qué resultados del programa examinar y rechaza tomar como punto de partida los objetivos de éste. Mantiene que, obrando así, el evaluador está más capacitado para identificar los verdaderos logros (y fracasos) del programa.

2.5. Evaluaciones basadas en la teoría

Las evaluaciones de metas, y con menor frecuencia las evaluaciones libres, pueden o no estar regidas por la teoría. Los modelos de evaluación basados en la teoría –que incluyen el modelo de teorías del cambio, así como la evaluación de la teoría del programa (Weiss, 1997; Rodgers *et al.,* 2000) y algunos aspectos de la evaluación realista (Pawson y Tilley, 1997)– no se distinguen por los resultados que tratan de explicar sino por su interés en descifrar la secuencia lógica o teórica por la que se espera que una intervención dé los resultados deseados. Localizando los factores determinantes o causales que se consideran importantes para el éxito, y cómo pueden interactuar, se puede decidir qué pasos deberían seguirse con mayor atención a medida que se desarrolla el programa (Rossi *et al.,* 1999).

Huey-Tsyh Chen, uno de los autores más influyentes en el desarrollo del concepto y la práctica de la evaluación basada en la teoría, ha argumentado que una desafortunada consecuencia de las evaluaciones no basadas en la teoría (como muchas de las evaluaciones que utilizan pruebas efectuadas al azar; véase más adelante) es que los resultados de la evaluación aportan visiones a menudo limitadas y a veces distorsionadas de los programas (Chen y Rossi, 1983). Las teorías que desea construir no son globales ni ambiciosas sino "modelos verosímiles y defendibles de cómo se puede esperar que funcionen los programas" (Chen y Rossi, 1983), en los cuales basar el ejercicio de evaluación en cualquier estadio de una intervención dada.

2.6. Evaluaciones formativas y conclusivas

2.6.1. Introducción

A veces se conoce a la evaluación formativa como evaluación del proceso y la evaluación conclusiva se denomina a veces evaluación del impacto. Las evaluaciones intermedias son generalmente formativas, aunque también pueden tener elementos sumativos; las evaluaciones ex-post –véase más arriba– suelen ser conclusivas, aunque también pueden examinar los procesos subyacentes a una intervención.

La evaluación formativa se pregunta cómo, por qué y en qué circunstancias funciona o no una política, un programa o un proyecto: estas preguntas son importantes a la hora de determinar la puesta en práctica eficaz de las políticas, los programas o los proyectos. Este tipo de evaluación suele buscar información en los factores, mecanismos y procesos contextuales que subyacen al éxito o el fracaso de una política. A menudo esto implica formular preguntas como para quién ha funcionado o no una política y por qué.

Las evaluaciones conclusivas formulan preguntas como qué impacto, si es que existe alguno, tiene una política, un programa o un proyecto sobre diversos grupos de gente: tiene por objetivo contrastar los efectos de una política –duros o blandos[2] (Lloyd y O'Sullivan, 2004)– con lo que se esperaba de ella en el estadio de su diseño, con otra intervención o con la ausencia de intervención alguna (contrafactual). La evaluación del impacto es por tanto la identificación sistemática de los efectos –positivos o negativos, deseados o no– causados por una determinada intervención. La evaluación del impacto ayuda a entender mejor hasta qué punto las actividades alcanzan a los grupos a los que están destinadas y la magnitud de sus efectos.

La distinción entre evaluaciones conclusivas y formativa, sin embargo, no es tan clara como se podría deducir de estas definiciones. Por ejemplo, los partidarios del modelo de teorías del cambio (Chen, 1990; Cornell y Kubisch, 1995; Funnel, 1997; Owen y Rodgers, 1999; Weiss, 1997) sostienen que para determinar si una política ha funcionado o no o si ha sido efectiva es imprescindible formular preguntas sobre cómo ha funcionado, para quién, por qué y bajo qué condiciones. No obstante, en la bibliografía sobre evaluación de políticas suele diferenciarse entre evaluar si una política ha sido efectiva (evaluación conclusiva) y evaluar por qué lo ha sido (evaluación formativa). Acontinuación nos centramos en las evaluaciones conclusivas o de impactos, el tipo más común de evaluación y en el que mayor diversidad metodológica puede encontrarse.

Los análisis de procesos e impactos tienden a seguir el siguiente modelo conceptual de intervenciones mediante políticas: las administraciones, las agencias o los operadores ponen en marcha medidas utilizando diversos medios o recursos, financieros, humanos, técnicos u organizativos. La inversión da lugar a una serie de resultados físicos –por ejemplo, kilómetros de carreteras construidos, número de centros de formación creados, etc.– que demuestran los progresos generados por la puesta en práctica de la medida. Los resultados son los efectos (inmediatos) en los beneficiarios directos de las acciones financiadas, como por ejemplo la reducción de la duración de los desplazamientos, los costes de transporte o la cantidad de personal formado. Estos resultados pueden expresarse según sus impactos a la hora de conseguir los objetivos globales o específicos del programa, y forman la base principal para valorar el éxito o el fracaso del servicio en cuestión. Los impactos específicos pueden incluir, por ejemplo, el incremento del tráfico de mercancías o una formación más acorde con las demandas

del mercado laboral. Los impactos globales están relacionados con la meta general de las ayudas, como la creación de empleo. Evidentemente, la medición de este tipo de impacto es compleja, y a menudo es difícil establecer relaciones causales claras (Comisión Europea, 1999).

Los principales métodos para estimar el impacto pertenecen a uno de estos grupos: métodos experimentales, que son esencialmente pruebas efectuadas al azar, y métodos cuasi-experimentales. Ambos intentan satisfacer la cuestión principal: estimar la adicionalidad de las intervenciones a partir de un cálculo de lo que hubiese sucedido de no haber existido el programa (contrafactual).

2.6.2. Métodos experimentales

La dificultad de la estimación contrafactual es evidente: en un momento dado se observa a un individuo, afectado o no por el programa. En la mayoría de los casos, comparar al mismo individuo a través del tiempo no nos proporcionará una estimación fiable del impacto que el programa pueda haber tenido sobre él, dado que desde que se introdujo el programa pueden haber cambiado para él muchas más cosas. Por tanto, no podemos pretender obtener una estimación del impacto del programa en cada individuo; lo más que podemos esperar es poder obtener el impacto medio del programa sobre un grupo de individuos comparándolo con un grupo similar que no haya estado expuesto al programa. Así pues, el objetivo crítico de la evaluación del impacto es establecer un *grupo de comparación*[3] creíble, un grupo de individuos que *en ausencia del programa* hubiesen experimentado fenómenos similares a los de aquellos que sí estuvieron expuestos al programa. Este grupo nos da una idea de qué le hubiese sucedido al grupo del programa si no hubiese estado expuesto al mismo, y nos permite obtener una estimación del impacto medio en el grupo en cuestión.

Se considera que los métodos de evaluación al azar, que implican la recopilación de información sobre el grupo del proyecto y el grupo de comparación en dos o más momentos, proporcionan el análisis más riguroso del impacto del proyecto y la contribución de otros factores. En la 'evaluación al azar antes y después del ensayo', o 'evaluación del ensayo en circunstancias de azar controlado', los sujetos –familias, escuelas, comunidades, etc.– son atribuidos al azar a los grupos de intervención y control. El azar no garantiza que ambos grupos vayan a ser idénticos, pero reduce la influencia de los factores extrínsecos, garantizando que las diferencias entre los dos grupos estarán libres de un sesgo sistemático. Los ensayos efectuados en circunstancias de azar controlado abordan el problema de que otros factores posibles influyan en el resultado exponiendo al grupo de experimentación y al de control a exactamente las mismas circunstancias, excepto la política, el proyecto o el programa que se está investigando; para una discusión clásica sobre este tema, véase Campbell y Stanley (1966).

Boruch (1997) argumenta que cualquier programa, sea social, educativo o de bienestar, debería ser estudiado de manera sistemática, empleando métodos experimentales en circunstancias de azar controlado para reunir pruebas válidas y fiables. Sin embargo, no está claro que se puedan evaluar todos los programas de este modo: por ejemplo, el examen de un tema como la independencia de un banco central tendría que sustentarse sobre otros métodos de evaluación. Los programas dirigidos a individuos o comunidades locales (como los servicios sanitarios, la reforma del gobierno local, la educación y la sanidad), en cambio, son mejores candidatos a evaluaciones aleatorias.

Lo que es más, los ensayos no permiten responder a todas las preguntas relacionadas con la evaluación: su problema principal es que dan una idea aproximada de la adicionalidad

neta, pero no proporcionan modo alguno de distinguir los números de gente para la que el programa mejora los efectos de los números de gente para quien los empeora. Este hecho limita seriamente el punto hasta el que se pueden identificar los beneficios concretos del programa sobre los individuos. También hay que tener en cuenta que los ensayos aleatorios sobre individuos son de poca utilidad si uno de los objetivos es producir un 'cambio de cultura' general que concierna a toda la población afectada: en esos casos es casi imposible evitar la contaminación del grupo de control. Para los programas de este tipo, la única opción factible para llevar a cabo un ensayo al azar puede ser la división aleatoria de áreas.

Aunque el diseño de ensayos realizados en circunstancias de azar controlado es atractivo por su simpleza, su ejecución puede ser compleja, y requiere una experiencia operativa y analítica considerable. Normalmente hay problemas éticos, suscitados por el hecho de exponer a un grupo de gente (el grupo experimental) a una política potencialmente nociva o, a la inversa, por el hecho de privar a otro grupo de gente (el grupo de control) de una política potencialmente beneficiosa. Sin embargo, en ausencia de pruebas sólidas *a priori* de que una política vaya a ser nociva o beneficiosa, suele creerse que es éticamente aceptable efectuar un ensayo para dirimir este extremo, siempre y cuando se interrumpa el ensayo en cuanto se hayan reunido pruebas válidas y fiables (Davies, 2004). Aun así, mucha gente opina que los métodos cuasi-experimentales descritos en lo que resta de este artículo proporcionan resultados razonablemente fiables y evitan algunos de los problemas (relacionados con la ética y los recursos) que suscitan los ensayos efectuados en circunstancias de azar controlado.

2.6.3. Métodos cuasi-experimentales

Si se descarta el modelo de ensayo aleatorio o se lo considera inapropiado, el resto para los evaluadores es elegir un método cuasi-experimental alternativo que pueda dar resultados razonablemente consistentes. Para esto hay una gama de métodos. En este artículo consideraremos:

- métodos de antes y después
- métodos de intervalos cronológicos
- mejoras de doble diferencia
- métodos de comparación de dos grupos equivalentes
- modelado estadístico de datos ya existentes para la evaluación de programas voluntarios
- métodos de comparación de áreas equivalentes[4]
- análisis económicos

2.6.3.1. Métodos de antes y después

En los métodos de 'antes y después' se identifica a la población a la que afectará una intervención antes y después de que se introduzca el programa. Se selecciona, de entre la población en cuestión, un 'grupo de programa' después de que se introduzca un programa y un 'grupo de comparación' antes de que se introduzca el programa (Greenberg y Morris, 2003). Después se recopilan los resultados de ambos grupos, y es la diferencia entre resultados lo que proporciona el cálculo aproximado de la adicionalidad. En este método es de la máxima importancia que en el periodo previo al programa se pueda establecer quiénes son los candidatos elegibles. Este modelo se ha utilizado frecuentemente en las evaluaciones, ajustando, por lo general, los resultados para controlar el efecto de las características observables.

Los indicadores de antes y después toman en consideracion la selección de individuos basada en características inobservables. La presuposición distintiva de este indicador es que

la diferencia entre el contrafactual verdadero posterior al programa y los resultados anteriores al programa da una media de cero en el conjunto de los individuos participantes en el programa. Principalmente, el indicador de antes y después presupone que las características inobservables son de dos tipos: las particulares de un individuo estables en el tiempo (efectos individuales) y las particulares de un individuo pero no estables en el tiempo (efectos transitorios). Se cree que la participación en el programa depende del efecto fijo y no del transitorio. Es una presuposición muy fuerte que puede verse violada, por ejemplo, por cambios macroeconómicos entre los dos puntos de observación.

2.6.3.2. Métodos de intervalos cronológicos

El problema de cómo deslindar los cambios introducidos por el programa y los cambios históricos en los estudios de antes y después puede afrontarse en ciertas ocasiones ampliando el número de periodos anteriores y, si es posible, el de periodos posteriores, de manera que se obtenga una serie de intervalos cronológicos. Si en esta serie se da una ruptura en el momento en que se introduce el nuevo programa o poco después, se interpreta que ése es el impacto del mismo. Los métodos basados en este principio se conocen como 'métodos de intervalos cronológicos interrumpidos'. Los intervalos cronológicos ayudan a descartar algunas de las posibles explicaciones de un cambio; en particular, siempre y cuando no se introduzcan programas relacionados al mismo tiempo, un repentino cambio en la serie resulta bastante concluyente. Esto es particularmente cierto cuando el cambio observado con el programa es mayor que el observado entre cualesquiera de los periodos anteriores. Si también puede probarse que el cambio que coincide con la introducción del programa perdura en el tiempo, las pruebas son más sólidas: éste es el motivo por el que interesa utilizar varios periodos posteriores.

Debido a la necesidad de series de datos razonablemente largas en un análisis de intervalos cronológicos interrumpidos, este método resulta más indicado para el análisis de datos administrativos, aunque en algunos casos pueden utilizarse estudios de gran escala continuos o casi continuos.

No obstante, los métodos de intervalos cronológicos interrumpidos no resuelven por completo el problema de la evaluación: si sucede que la introducción del programa en cuestión coincide con otros eventos, o con la introducción de otros programas que tengan un impacto en los resultados, será imposible probar que el programa en cuestión ha causado el cambio que se ha observado. En el caso de los programas que tienen un impacto retardado o gradual, la interrupción en la serie de periodos tendrá lugar un tiempo después de que se introduzca el programa. A no ser que esto se tenga en cuenta de antemano, puede ocurrir que el impacto del programa pase desapercibido.

Igual que pasa con el método básico de antes y después, los modelos basados en intervalos cronológicos interrumpidos resultan más convincentes en el caso de programas con un impacto razonablemente grande, pues será relativamente fácil detectar dicho impacto entre el 'ruido ambiente'. Esto significa que este método no sería adecuado para programas voluntarios, especialmente aquellos con un bajo número de participantes.

Un tipo particular de método de intervalos cronológicos interrumpidos es el método de 'tratamiento suprimido'. En circunstancias excepcionales se introduce un programa que después se suprime. En esos casos, se esperaría que las series cronológicas revelasen dos 'interrupciones': una en el momento en que se introduce el programa y otra cuando se lo suprime. Se

esperaría que en gran medida el primer cambio quedase revertido tras la supresión del programa. En principio, este modelo podría ser muy pujante; sin embargo, se usa pocas veces o ninguna en las evaluaciones de los programas gubernamentales por la simple razón de que cuando se abandona una política no suele haber demasiado interés en practicarle la autopsia.

En último lugar, en los métodos de tratamiento demorado se ensaya un programa en unas pocas zonas antes de introducirlo en la nación entera. Si los datos recabados en los intervalos cronológicos muestran una diferencia entre antes y después en las zonas de ensayo, y después esa diferencia se reproduce en otras zonas cuando se implanta el programa a escala nacional, se obtienen pruebas muy firmes del impacto del programa.

Cuando se trabaja con métodos de intervalos cronológicos interrumpidos es importante asegurarse de que las variables que se están midiendo permanezcan estables en el tiempo: cuando las definiciones y las prácticas de cómputo varían a menudo en el tiempo, es mucho más difícil –y a veces imposible– utilizar esos datos como medidas válidas o atribuir un motivo a una interrupción en la serie de intervalos.

2.6.3.3. Mejoras de doble diferencia

Tanto el método básico de antes y después como el método de los intervalos cronológicos interrumpidos computan la adicionalidad atribuible a un programa calculando la diferencia (ajustada) en los resultados entre los periodos anteriores y posteriores. Con todo, ambos métodos pueden ser poco convincentes si se da la sospecha de que la diferencia puede deberse a otros eventos que tuviesen lugar más o menos al mismo tiempo que se implantó el programa. Este problema puede afrontarse mediante los métodos de doble diferencia.

Este tipo de método funciona comparando una estimación de antes y después referente a los participantes con una estimación de antes y después referente a los no participantes, y entendiendo la diferencia como adicionalidad. La presuposición de identificación es más verosímil que en el caso del indicador de antes y después. Específicamente, se parte de la base de que la media del cambio en la medición del resultado ajeno al programa será la misma para los participantes y los no participantes. En la práctica, esto significa que el indicador de doble diferencia puede asimilar los cambios macroeconómicos o los cambios en el ciclo vital, siempre y cuando éstos afecten de manera similar a participantes y no participantes. Esto reafirma la necesidad de seleccionar a un grupo adecuado de no participantes. A menudo la elección del grupo de comparación se justifica porque tiene tendencias similares a las del grupo experimental, en lo que respecta a la variable en los resultados en cuestión, a lo largo de un periodo prolongado antes de la implantación del programa. Aunque esto resulta tranquilizador, hay que tener en cuenta que es habitual ajustar las estimaciones de doble diferencia para destacar las características observables; por lo tanto, la tendencia que hay que observar es la de los resultados regresivos, no la de las propias medidas de resultados.

Además de los efectos individuales y transitorios que caracterizan al indicador de antes y después, también se tiene en cuenta un efecto común a los individuos pero que varía en el tiempo (efecto de tendencia). Como ya se ha hecho notar, el indicador de antes y después elimina los efectos individuales. La ventaja del indicador de doble diferencia es que también elimina los efectos de tendencia. Así pues, el único efecto que permanece es el específico del individuo que varía en el tiempo: éste no puede ser controlado y, si influenciase la decisión de participar en el programa, la presuposición identificativa quedaría en entredicho y las estimaciones resultadas estarían sesgadas (Heckman y Smith, 1999).

Lo que es más, tanto el indicador de antes y después como el de doble diferencia dependen de que la composición del grupo experimental no sufra cambios en el periodo posterior al programa. Si esto no se cumple, la diferencia entre la verdadera hipótesis y el resultado previo al programa no dará necesariamente una media de cero en todos los individuos. Un cambio así en la composición es más probable en la repetición de los datos de un corte transversal, pero también puede darse con datos longitudinales si la muestra fuese reduciéndose con el paso del tiempo de manera sistemática.

Además de los grupos de control no equivalentes sobre los que se ha hablado antes, es posible utilizar variables no equivalentes de resultados como modelos de doble diferencia. Este método compara el cambio en el tiempo en los resultados bajo examen con los cambios en otros resultados relacionados. Estas variables no equivalentes de resultados tienen que satisfacer el requisito de haber cambiado históricamente en paralelo a los resultados de interés, pero no les afecta la política que se está examinando.

Un ejemplo famoso de este método fue la evaluación de las pruebas obligatorias de alcoholemia en el Reino Unido, en 1967. Se compararon dos variables de resultados: los accidentes en las noches del fin de semana, cuya frecuencia se esperaba que cambiase con la implantación de las pruebas, y los accidentes en las horas puntas laborales, en cuya frecuencia no se preveían cambios. Lo que se observó fue una brusca caída en la primera variable y pocos o ningún cambio en la segunda: esto dio pruebas sólidas de que las pruebas habían tenido un impacto. De nuevo, la principal razón por la que no suele utilizarse este método es la dificultad de encontrar resultados pertinentes con los cuales comparar la principal variable de resultados.

2.6.3.4. Métodos de comparación de dos grupos equivalentes

Los modelos de evaluación basados en los intervalos cronológicos son menos convincentes en el caso de programas voluntarios con bajo número de participantes. Por este motivo se han diseñado métodos cuasi-experimentales que funcionan precisamente en esas condiciones, y pueden además afrontar algunas de las limitaciones de los métodos de doble diferencia. Estos métodos son fundamentalmente dos: el método de comparación de dos grupos equivalentes que se describe aquí y otro modelo más general que puede ser denominado 'modelado estadístico de datos ya existentes', dado que engloba varios métodos de análisis en vez de ser un modelo alternativo cuasi-experimental como tal. El primero de estos dos modelos se utiliza normalmente cuando hay que recopilar datos sobre los resultados por medio de un sondeo, caso en el cual el tamaño de los grupos de participantes y no participantes tendrá que ser razonablemente pequeño. Los métodos de 'modelado estadístico' se utilizan cuando hay disponibles datos sobre los resultados, procedentes por lo general de los registros administrativos o de sondeos elaborados para otro propósito.

El método de comparación de dos grupos equivalentes hace uso del hecho de que, en los programas voluntarios, la adicionalidad se concentra en los participantes en el programa en vez de estar distribuida por toda la población potencialmente afectada. De hecho, se da por sentado que el programa no tiene impacto alguno sobre los no participantes, de manera que de este grupo se puede extraer una estimación hipotética razonable de cuáles serían los resultados en ausencia del programa. Igual que con todos los métodos cuasi-experimentales, se comparan dos grupos: un grupo de programa y un grupo de comparación. En el método de comparación de dos grupos equivalentes, el grupo de programa se selecciona de entre los participantes y el grupo de comparación de entre los no participantes, y se compilan datos

acerca de los resultados de los miembros de los dos grupos. Si se considera que los datos de resultados son caros de recopilar –lo cual sucedería si hubiese que hacerlo por medio de un sondeo–, se escogerá un grupo de comparación del mismo tamaño que el grupo de programa (de ahí que la comparación sea entre dos grupos *equivalentes*), aunque también se pueden comparar grupos no equivalentes.

El método de la equivalencia da por sentado que la selección puede explicarse puramente mediante las características observables. En principio, la aplicación del método es sencilla: por cada individuo del grupo de programa se busca un individuo equivalente en el grupo de comparación. La elección de este emparejamiento viene dictada por las características observables. Lo que se necesita es casar a cada individuo del grupo de programa con un individuo de características similares; se puede entonces calcular el efecto medio del tratamiento como la diferencia media en resultados entre un grupo y otro.

Los métodos de comparación de dos grupos equivalentes son controvertidos por la forma de selección del grupo de comparación. Para que un grupo de comparación proporcione una estimación hipotética ajustada de las condiciones caso de no existir el programa, el grupo debe tener el mismo perfil que el grupo de programa en lo referente a los resultados, con la única excepción de que no está expuesto al programa (o, en este caso, *decide* no estarlo). Pero, al contrario que en el caso de los programas obligatorios, en este caso el grupo de programa se escoge a sí mismo y, en gran medida, sus razones para participar no se conocerán o comprenderán por completo; en consecuencia, resulta muy difícil construir un grupo de comparación adecuado.

Así, para evitar un sesgo en el método de comparación de dos grupos equivalentes, deben satisfacerse los dos criterios que siguen:

- los evaluadores deben conocer exhaustivamente los factores que influyen en la participación en una política
- deben existir datos sobre estos factores para todos los participantes y no participantes, o para una parte de ambos

Si estos dos criterios se cumplen, la comparación entre dos grupos equivalentes es un método sólido de evaluación; si no se cumplen, la equivalencia no será perfecta y se darán diferencias residuales e incontroladas entre el grupo de programa y el de comparación. Estas diferencias introducirán un sesgo en las estimaciones de adicionalidad.

2.6.3.5. Modelado estadístico de datos ya existentes

El método de comparación de dos grupos equivalentes crea un grupo de comparación de más o menos el mismo tamaño que el grupo de programa, pero que se selecciona de entre un colectivo mucho mayor: el de no participantes. Esto es adecuado si la recopilación de datos es cara, dado que en ese caso la selección de un grupo de comparación mayor de lo estrictamente necesario es un despilfarro. Sin embargo, si pueden recopilarse datos con un coste bajo o sin coste alguno, se puede obtener una mayor precisión usando un grupo mayor de no participantes. La bibliografía sobre econometría sugiere varias maneras de estimar hipotéticamente los efectos de la no existencia del programa bajo estas circunstancias, maneras que son todas de naturaleza técnica e incluyen: el cotejo de puntuaciones de propensión por medio de evaluación *kernel,* el indicador de variables instrumentales y el indicador de selección Heckman. Todos estos métodos necesitan presuposiciones muy importantes para una estimación hipotética objetiva y tienden a caer en los mismos problemas de interpretación que el método estándar de comparación de dos grupos equivalentes.

2.6.3.6. Métodos de comparación de áreas equivalentes

La comparación de áreas equivalentes sólo puede utilizarse para medir la adicionalidad cuando se ensaya una política en varias zonas geográficas: la forma básica de la comparación de áreas equivalentes parte de un pequeño número de zonas en las cuales se prueba la nueva intervención, de entre las cuales se selecciona al grupo de programa. Después se emparejan las zonas con otras áreas de características similares donde no se esté probando la intervención y se selecciona de entre ellas al grupo de comparación. En una variante del método básico, se selecciona el grupo de comparación de entre el resto del país, no de zonas equivalentes. A continuación se recopilan los resultados de ambos grupos.

La interpretación de los métodos de comparación de áreas equivalentes puede resultar problemática, porque, si bien las diferencias observadas entre el grupo de programa y el de intervención pueden atribuirse a la intervención de la política, también podrían achacarse a diferencias en el perfil de los dos grupos o a otras características locales que no se hubiesen tenido en cuenta durante el proceso de emparejamiento. La primera diferencia es la que deseamos medir; las otras son un problema potencial, pero hasta cierto punto pueden abordarse buscando diferencias entre los grupos en el estadio de análisis, por ejemplo, por medio de análisis regresivos o midiendo las muestras con respecto a un perfil común.

Las comparaciones de áreas equivalentes son candidatos naturales a un modelo basado en diferencias y similitudes. Si pueden recopilarse datos de varios intervalos cronológicos sobre los resultados para la zona piloto y la de control y se puede probar que estas dos series de intervalos habían discurrido en paralelo, durante al menos un tiempo, antes de que se introdujese el programa en las zonas piloto, tomar la diferencia entre antes y después en las dos zonas y compararlas puede proporcionar un indicador del efecto del programa mucho mejor que el que se obtendría mediante la simple diferencia entre las dos zonas. Este modelo es especialmente útil si el emparejamiento de las zonas es meramente aproximado, dado que en ese caso puede haber diferencias previas importantes entre ambas zonas.

2.7. Análisis económicos

Los análisis económicos proveen una visión particular de la evaluación. Estos métodos relacionan el impacto con otros estadios de un ciclo de intervención, principalmente las aportaciones y el objetivo de la política, para medir la eficacia y la eficiencia. La eficacia compara lo que se ha hecho con lo que estaba previsto originalmente, es decir, compara los resultados y/o impactos reales con los que se esperaban, mientras que la eficiencia examina la relación entre los resultados y/o impactos con los recursos (principalmente los recursos financieros) empleados para conseguirlos. Los análisis económicos incluyen los sondeos de control del gasto público y técnicas más sofisticadas, como análisis coste - beneficio, coste - eficacia y coste - utilidad.

2.7.1. Sondeos de control del gasto público

Los sondeos de control del gasto público siguen el flujo de fondos públicos y determinan hasta qué punto los recursos llegan realmente a los grupos a los que están destinados. Los sondeos examinan la manera, la cantidad y el tiempo en que se conceden recursos a los diversos niveles del gobierno, especialmente a las unidades responsables de la prestación de servicios sociales como la sanidad o la educación. Los sondeos de control del gasto público pueden simplemente comparar los costes y gastos programados de diferentes iniciativas sin

tener en cuenta los resultados a conseguir o conseguidos. Las limitaciones de semejantes valoraciones y evaluaciones resultan evidentes –puesto que nos indican muy poco sobre la *eficacia relativa o los efectos beneficiosos* de las diversas intervenciones– y, de por sí, tienen muy poco valor para la evaluación de políticas. Sin embargo, muchas veces estos sondeos se llevan a cabo como parte de unos sondeos más amplios sobre la prestación de servicios, que se centran en la calidad del servicio, las características de su prestación, su gestión o sus estructuras incentivas, y que contribuyen a fomentar la responsabilidad cuando hay poca información financiera disponible.

2.7.2. Análisis de la relación entre coste y beneficio, entre coste y eficacia y entre coste y utilidad

Otros tipos de valoración y evaluación económica, más pujantes y más útiles para la creación de políticas, incluyen los análisis de las relaciones coste-eficacia y coste-beneficio, herramientas para juzgar si los resultados e impactos de una actividad pueden justificar sus costes. El análisis de la relación coste-beneficio mide tanto las aportaciones como los resultados en términos monetarios (World Bank, 2004). El análisis de la relación coste-eficacia mide las aportaciones en términos monetarios y los resultados en términos cuantitativos no monetarios (como las mejoras en la alfabetización de los escolares). Ambas técnicas pueden contribuir a saber qué proyectos dan más resultado a cambio de la inversión, y también han demostrado ser herramientas útiles para convencer a quienes diseñan las políticas y controlan los recursos de que los beneficios de una intervención la justifican, aunque en la práctica es frecuente que no se disponga de datos necesarios para los cálculos de la relación coste-beneficio y los resultados previstos dependan en gran medida de las presuposiciones adoptadas. El análisis coste-beneficio suele implicar la consideración de usos alternativos de un recurso determinado o el *coste en oportunidades* de hacer algo frente a hacer otra cosa. Otro tipo de valoración económica es el análisis de la relación entre coste y utilidad, que evalúa la utilidad de diversos resultados para diversos usuarios o consumidores de una política o servicio. Normalmente, este método implica valoraciones y evaluaciones subjetivas de resultados utilizando datos cualitativos y cuantitativos.

Estos análisis usan diversas herramientas para estimar los costes y los beneficios de las políticas a lo largo del tiempo, como la *tasa de descuento* para ajustar el valor de los resultados que tendrán lugar en el futuro.

2.8. Evaluaciones cuantitativas y cualitativas

Una última distinción dentro de los tipos de evaluación se refiere al carácter de los datos recogidos y la manera en que se analizan. En este sentido es habitual distinguir entre evaluaciones cuantitativas y cualitativas. Las evaluaciones cualitativas están pensadas para "permitir al evaluador estudiar determinados temas en profundidad y detalle" (Patton, 2002). Normalmente la profundidad y el detalle son imprescindibles para decidir qué preguntas formular en una evaluación y para identificar las condiciones situacionales y contextuales bajo las cuales funciona o deja de funcionar una política, un programa o un proyecto. Los métodos cualitativos de evaluación son especialmente importantes para las evaluaciones formativas, las cuales, como apunta una vez más Patton (2002), suelen limitarse por completo a un contexto específico. Sucede a menudo que en las evaluaciones cualitativas no hay intentos de generalizar los hallazgos más allá del contexto en que uno está trabajando.

En cambio, las evaluaciones cuantitativas pueden usar técnicas econométricas como las ya mencionadas en este capítulo para medir el impacto o pueden resumir grandes cantidades de datos compilados con sondeos hechos ex profeso.

Elliot Eisner ha argumentado que, aunque las técnicas cuantitativas pueden proporcionar alguna información útil, "la evaluación requiere un mapa interpretativo sofisticado, no sólo para separar lo trivial de lo significativo sino también para entender el significado de aquello que se conoce" (Eisner, 1994). Sin embargo, en última instancia, la decisión de adoptar métodos cuantitativos o cualitativos depende de los fines de la evaluación y de la naturaleza de la intervención que se evalúa, aunque también tienen importancia las tendencias generales del clima político. En el contexto de la evaluación de intervenciones de desarrollo, White (2005) apunta que el abandono del crecimiento como medida de desarrollo en la década de 1970 se reflejó en una modificación de las maneras de calcular la eficacia. Esta modificación obedeció en parte a la opinión de que los sectores sociales eran menos receptivos a los análisis económicos de la relación entre coste y beneficio y a un deseo de centrarse directamente en resultados no económicos, como la igualdad de género. Se pensaba que el análisis de la relación entre coste y beneficio no podía aprehender estos temas y que, por tanto, se hacía necesario un modelo más cualitativo. Para la década de 1980, los modelos cualitativos habían pasado a dominar los estudios de evaluación que se hacían para las agencias de desarrollo, cambio que en ese momento se vio reforzado por la importancia que se daba al proceso.

Sin duda éstos son puntos importantes que pueden pasarse por alto en un estudio estrictamente económico; y suele ser justo decir que los 'proyectos de proceso', cuyo fin principal suele ser el desarrollo institucional, están habitualmente demasiado distantes de los resultados finales de desarrollo como para que pueda cuantificarse su impacto en estos últimos. No obstante, la nueva centralidad concedida a los resultados en el contexto de iniciativas como los objetivos de desarrollo del milenio de las Naciones está haciendo que los creadores de las políticas estén volviendo a prestar atención a las evaluaciones cuantitativas. Los métodos cualitativos tienen dificultades para responder preguntas cruciales para los responsables de las políticas, como: ¿hasta qué punto están las intervenciones de las agencias trayendo progreso en los campos relacionados con las metas de desarrollo del milenio?

En la próxima sección nos centraremos en los métodos de compilación de datos.

3. MÉTODOS DE COMPILACIÓN DE DATOS

3.1. Introducción

Los diferentes tipos de evaluación que se han presentado pueden acometerse por medio del uso de varios métodos de compilación de datos, que se presentan en esta sección. Aunque algunas evaluaciones utilizan un único método para la compilación de datos, muchas utilizan una mezcla de métodos, combinando los datos de varios instrumentos y enfoques. En esos casos, se pone especial cuidado en evitar que los datos se pierdan o dupliquen como consecuencia de tener que combinar datos de diferentes fuentes. Los métodos primarios y secundarios que se presentan aquí son:

- informe documental
- sondeo formal
- entrevistas
- grupos de discusión y otras formas de consulta

- monografías
- observación de participantes
- métodos participativos[5]

3.2. Revisión documental

La mayor parte de las evaluaciones tienden a hacer uso de la revisión de documentos. En algunos casos, los documentos son la fuente de 'reconstrucción' de una base para la evaluación, dado que registran lo que un proyecto pensaba conseguir, las condiciones desde las cuales se desarrolló, cómo se desarrolló, qué cambios se llevaron a cabo y cómo se llevaron a la práctica. En otros casos, el análisis documental acompaña a otros métodos de compilación de datos, como por ejemplo entrevistas en profundidad o sondeos.

Las revisiones documentales tienen grandes ventajas: además de la precisión añadida que proporcionan los datos documentales y administrativos, está también la ventaja de que los datos no suelen estar sujetos a los sesgos que sufren los sondeos debido a los casos en que no hay respuesta, y el coste de la compilación de datos es menor. Sin embargo, el evaluador tiene menos libertad para determinar el contenido de los datos a su disposición.

Cada vez se utilizan más las revisiones sistemáticas de la bibliografía de investigación existente como método válido y fiable de utilizar la evidencia generada por la investigación; también pueden permitir el establecimiento de una visión acumulativa de dicha evidencia. Las revisiones sistemáticas difieren de otros tipos de síntesis de la investigación (por ejemplo, de las revisiones narrativas y el 'recuento de votos') en que son más rigurosas a la hora de buscar y encontrar evidencias, dado que tienen criterios explícitos y transparentes para valorar la calidad de la evidencia generada por la investigación, y especialmente para identificar y controlar los diversos tipos de sesgo en los estudios existentes, en cuanto tienen maneras explícitas de establecer la posibilidad o imposibilidad de comparar distintos estudios y, en consecuencia, de combinar y establecer una visión acumulativa de lo que nos dice la evidencia actualmente existente.

Dos métodos comunes de sintetizar la evidencia actualmente existente son las revisiones narrativas y el 'recuento de votos'. Las revisiones narrativas tratan de identificar la bibliografía disponible acerca de un tema, las metodologías utilizadas, las conclusiones alcanzadas hallazgos y sus salvedades. Pueden presentar un resumen de la investigación sobre un tema o no. Tienden a identificar el alcance y la diversidad de la bibliografía disponible, de la cual mucha será inconsistente o no concluyente. Una de sus grandes limitaciones es que las revisones narrativas son casi siempre selectivas. No siempre implican una búsqueda sistemática de toda la bibliografía relevante usando fuentes electrónicas e impresas, además de pesquisando directamente para encontrar estudios inéditos o trabajos en curso de elaboración: esto significa que las revisones narrativas tradicionales sobre la bibliografía sufren a menudo de sesgos de selección o publicación. Las revisiones sistemáticas difieren también de las narrativas en que explicitan los criterios de búsqueda para identificar la bibliografía disponible y los procedimientos por los que se valora e interpreta críticamente dicha bibliografía. Esto proporciona un grado de transparencia por medio del cual los lectores pueden determinar qué evidencia se ha estudiado y cómo se ha interpretado y presentado.

Los recuentos de votos tratan de acumular los resultados de un conjunto de estudios relevantes contando "cuántos resultados son estadísticamente significativos en un sentido, cuántos son neutrales (es decir, que 'no tienen efecto') y cuántos son estadísticamente significati-

vos en el otro sentido" (Cook *et al.,* 1992: 4). Se considera que la categoría más numerosa, es decir, que tenga más votos, es la que representa los hallazgos típicos o modales, indicando pues el medio de intervención más eficaz. Un problema obvio de los recuentos de votos es que no tienen en cuenta el hecho de que hay estudios que son metodológicamente superiores a otros y, por tanto, merecen tener más peso. Las revisones sistemáticas de la literatura distinguen entre estudios que abordan muestras de mayor o menor tamaño, estudios de mayor o menor poder y precisión, y los valoran como corresponda.

Un tipo especial de revisión sistemática es el 'metaanálisis', que agrega los resultados de los estudios comparables y "combina los efectos de tratamiento calculado por cada estudio en un efecto de tratamiento del conjunto de todos los estudios" (Morton, 1999). Esto, sin embargo, no siempre es posible, puesto que sólo puede ocurrir si hay una consistencia real entre los estudios primarios en cuanto a los tipos de intervención que analizan, la población que estudian y los resultados que valoran. (Véase Deeks, Altman y Bradburn, 2001).

Slavin (1984, 1986) ha aventurado que el método que se use para generar evidencia investigadora es menos importante que la calidad de los estudios primarios emprendidos, usen las opciones metodológicas que usen. Sugiere que lo que se necesita es la 'síntesis de la mejor evidencia', en la que "los evaluadores apliquen criterios de inclusión consistentes, justificados y claramente expuestos *a priori"*, a los trabajos a examinar. Para Slavin, los estudios primarios deben ser "relevantes para el tema en cuestión, deben estar basados en un esquema que minimice el sesgo y tener validez externa".

Generalmente los resultados de las revisiones documentales se analizan por medio del análisis teórico (donde se estudian los documentos para saber si obedecen a una teoría o explicación predefinida), el análisis estructural (donde se estudia en detalle la estructura del documento, es decir, cómo está construido, en qué contexto se sitúa, cómo transmite sus mensajes) y el análisis de contenidos (donde se estudia y compara la información de determinados tipos de documento).

3.3. Sondeos formales

Los sondeos formales son un método para compilar información estandarizada a partir de una muestra seleccionada de individuos y organizaciones. A menudo, los sondeos compilan información comparable para un número de casos relativamente grande y proporcionan datos de base con los cuales comparar el rendimiento de una estrategia, un proyecto o un programa. Así pues, los sondeos pueden ser una fuente valiosa para una evaluación formal del impacto de un programa o proyecto. Hay diversos tipos de instrumentos para los sondeos que pueden utilizarse para recopilar la información necesaria para responder las preguntas de la investigación; entre éstos, los siguientes:

3.3.1. Cuestionarios

Los cuestionarios compilan información por medio de preguntas prefijadas. El cuestionario puede formularlo un entrevistador (cara a cara o por teléfono) o puede completarlo el entrevistado (en el caso de sondeos postales u *on line*). Los cuestionarios pueden recoger información fáctica e información referente a los comportamientos y las actitudes, además de medir los conocimientos de los entrevistados, aunque estos últimos sólo pueden recogerse fiablemente si un entrevistador formula el cuestionario o si el entrevistado lo rellena en un entorno controlado. La forma en que se recopilen los datos puede influenciar la fiabilidad y

precisión de la información obtenida. Por ejemplo, la precisión de la información acerca de comportamientos 'problemáticos', como el consumo de drogas, puede diferir dependiendo de si los datos los recoge un entrevistador o si el entrevistado rellenó un formulario.

3.3.2. Diarios

Los diarios posibilitan la compilación prospectiva de información, es decir, a medida que se produce un evento. Son una forma de sondeo a rellenar por el entrevistado, de quien se requiere que registre detalles del comportamiento de interés durante un periodo específico. Se espera capturar así los detalles del comportamiento habitual de los entrevistados. Los diarios pueden capturar información sobre el comportamiento mucho más detallada de lo que con frecuencia es posible en otros tipos de sondeo, y puede usarse al tiempo que los cuestionarios estructurados.

3.3.3. Mediciones

Pueden efectuarse mediciones para recopilar información fáctica como la estatura de los entrevistados, su peso, su tensión vascular, sus niveles de hierro en sangre, etc. Igual que en el caso de los diarios, estas mediciones pueden efectuarse en conjunción con la información obtenida de un cuestionario (o un diario). Es necesario desarrollar protocolos para asegurarse de que las mediciones se efectúan de manera estandarizada. Puede ser necesaria la aprobación ética.

3.3.4. Pruebas

Como parte del proceso de entrevistas del sondeo, pueden administrarse pruebas para medir la capacidad de los entrevistados para desempeñar determinadas tareas, como leer o caminar. Con frecuencia estas pruebas son herramientas estandarizadas de valoración que se han desarrollado para un contexto determinado, como la valoración clínica o educacional en un hospital o una escuela. Como en el caso de la compilación de mediciones, es necesario desarrollar protocolos que garanticen que las pruebas se administran de manera coherente y que puedan ser formulados (fiablemente) durante una entrevista correspondiente a un sondeo.

3.3.5. Observaciones

Pueden llevarse a cabo observaciones de información fáctica, como el estado de la vivienda del entrevistado. Los observadores deben haber recibido una cuidadosa formación para registrar la información de manera coherente. Los datos observacionales pueden compilarse al mismo tiempo que otros tipos de información para obtener una imagen más detallada de las circunstancias del entrevistado. La elección de instrumento de compilación de datos se verá influenciada por la naturaleza de las preguntas, el tipo de información que se requiere, el grado de detalle necesario, el grado requerido de precisión de los datos, las características de la población de la que se recaba la información, el tiempo y el dinero.

Los sondeos pueden ser transversales, si recopilan información sobre la población concernida en un momento determinado, o longitudinales, si reúnen información sobre individuos determinados a lo largo del tiempo. Además, los sondeos transversales pueden dividirse en sondeos *ad hoc* –hechos una sola vez–, sondeos continuos –en los que el trabajo de campo tiene lugar, por ejemplo, cada mes del año, siendo el muestreo de cada mes representativo de la población concernida– y sondeos repetidos –que tienen lugar en momentos regulares y determinados, como por ejemplo cada año o cada dos años, con el trabajo de campo concentrado en unos cuantos meses, lo cual permite medir el cambio global en el nivel agregado, dado que las estimaciones de un sondeo pueden compararse con otras de la misma serie

pero no permite saber si el cambio observado tuvo lugar gradualmente o no, cosa que sí permiten los sondeos continuos–. Igual que con los sondeos continuos, no hay nada en el modelo de los sondeos repetidos que requiera un solapamiento de las muestras en diferentes momentos. Esto los distingue de otros tipos de sondeo, como los de muestreo rotatorio o los estudios longitudinales sin rotación. Las encuestas de muestreo rotatorio se programan a intervalos regulares o continuamente y utilizan muestreos rotatorios, es decir, que se incluye a cierto número de gente en el sondeo, se le examina algunas veces y después se le excluye del sondeo: no se intenta seguir a los entrevistados o unidades de muestreo ni tampoco relacionar los resultados de unos u otros a través del tiempo para obtener estimaciones longitudinales. En los estudios longitudinales sin rotación, se sigue a través del tiempo a los participantes para crear un registro longitudinal; sin embargo, estos datos diacrónicos no pueden extrapolarse a la población general.

3.4. Entrevistas

Las entrevistas en profundidad son probablemente la forma más habitual de investigación cualitativa en la evaluación. Se cree que los informes personales orales tienen una importancia central en la investigación social debido a su poder de elucidar el significado (Hammersley y Atkinson, 1995). Los informes individuales y personales presentan el lenguaje que la gente utiliza y los aspectos que destacan, y permiten que la gente dé explicaciones claras acerca de sus acciones y decisiones.

Las entrevistas en profundidad dan la oportunidad de compilar datos ricos y detallados, dado que el entrevistador puede 'exprimir' el tema e incitar al entrevistado a que profundice más y más en sus respuestas. Son perfectas para la exploración en profundidad de un tema que proporcione al investigador una visión detallada del mundo del entrevistado, sus creencias, experiencias y sentimientos, y las explicaciones que da de sus convicciones o acciones. Estas entrevistas también se prestan bien a explorar procesos complejos o desentrañar la toma de decisiones. Un buen entrevistador establece una comunicación con su entrevistado, lo cual facilita la exploración de temas conflictivos, dolorosos o problemáticos.

El grado de estructuración de las entrevistas varía: un rasgo clave de las entrevistas cualitativas es que las preguntas no están determinadas de antemano, sino que el entrevistado tiene cierta influencia sobre la dirección y la cobertura de la entrevista. En algunos estudios, acaso particularmente en aquellos cuyo propósito es revelar al investigador un mundo social con el que no está familiarizado, la entrevista puede estar relativamente desestructurada, de manera que el entrevistador formula preguntas muy amplias y el entrevistado moldea el resultado. En otros casos, el investigador tendrá una consciencia más clara de los temas a explorar y jugará un papel más activo en la dirección de los temas a cubrir por la entrevista. A veces, los términos 'desestructurado' y 'semiestructurado' denotan diversos grados en los que la investigación dirige la entrevista, aunque no siempre se utilizan de manera coherente.

En los modelos biográficos suele usarse un tipo especial de entrevista, historias vitales y narraciones donde los investigadores retornan con frecuencia a un informante para obtener más datos o mantener más entrevistas. Por ejemplo, pueden estudiar la perspectiva de una familia entrevistando a diversos miembros de la misma, o la de los miembros de una comunidad específica.

3.5. Grupos de discusión y otras formas de consulta

Los grupos de discusión, o discusiones de grupo, consisten habitualmente en hasta diez personas reunidas para hablar de un tema o de varios. El grupo lo modera o facilita un investigador. Aunque los grupos de interés han adquirido una imagen un tanto turbia, son un método riguroso y bien establecido de investigación y evaluación social. En los grupos de discusión, los datos son moldeados y matizados por medio de la interacción del grupo: escuchar la participación de otros estimula el pensamiento y anima a la gente a reflexionar sobre sus propias opiniones o su comportamiento, generando nuevo material.

Los grupos de discusión pueden funcionar muy bien cuando se trata de temas abstractos o conceptuales, que en una entrevista podrían dejar en blanco al entrevistado. También pueden ser utilizados para temas conflictivos, siempre y cuando las características sociales de los participantes y su conexión con el tema de la investigación sean lo bastante similares como para crear un entorno que transmita sensación de seguridad.

Otras formas de consulta más innovadoras son las siguientes:

- El método Delphi (Adler y Ziglio, 1996; Cantrill *et al.,* 1996; Crichter y Gladstone, 1998). Éste es un proceso iterativo especialmente orientado a la predicción en el cual se pide a un grupo de expertos que respondan individualmente a una serie de preguntas, bien por medio de una encuesta bien usando investigación cualitativa. A continuación se difunden las respuestas entre los miembros del panel, a quienes se les pide que valoren sus propias respuestas, que luego se difunden y matizan sucesivamente hasta llegar a un consenso o una disensión acordada. El grupo no se reúne físicamente.
- La técnica de grupo nominal es una variante del método Delphi que sigue un patrón similar para obtener las primeras respuestas de los miembros del panel. Sin embargo, tras esa primera fase, se llevan a cabo nuevas iteraciones usando un formato un tanto similar a un grupo de discusión. El objetivo del grupo es alcanzar un consenso en las áreas de acuerdo y desacuerdo.
- Los jurados ciudadanos (Coote y Lenaghan, 1997; Davies *et al.,* 1998; White *et al.,* 1999). Se reúne a un grupo de entre doce y veinte personas durante varios días para escuchar a varios 'testigos' expertos y formularles preguntas, deliberar y debatir entre ellos y hacer recomendaciones sobre las acciones a tomar, que pueden estar consensuadas o no.
- Sondeos deliberativos (Fishkin, 1995; Park *et al.,* 1999). Se trata de un conjunto de actividades dirigidas a explorar cómo cambia la opinión pública cuando el público tiene la oportunidad de informarse en profundidad sobre un tema. Se lleva a cabo una encuesta para establecer un punto de referencia de la opinión pública. Los participantes asisten a un evento conjunto, por lo general durante un fin de semana, que incluye debates en grupos pequeños, conferencias de expertos y sesiones políticas en las que portavoces de los partidos responden a preguntas. Se repite la encuesta para medir en qué sentido y en qué medida ha cambiado la opinión.
- Congresos o talleres de consenso (Seargeant y Steele, 1998). En este modelo, un panel de unas quince personas trata de definir las cuestiones que desea afrontar con respecto a un tema determinado, consulta a expertos, recibe información, delibera y trata de alcanzar un consenso. El panel elabora su propio informe y lo presenta en un congreso abierto, donde se debate de nuevo.

- 'Valoración participativa'. Históricamente, este método se ha usado en el trabajo para el desarrollo de las colonias, pero ahora es más frecuente en el Reino Unido. Está pensado para involucrar a la gente, especialmente la de comunidades socialmente excluidas, en decisiones que atañen a sus vidas. Combina varias herramientas visuales (mapas) con discusiones en grupo y entrevistas semiestructuradas.
- 'Planificación de verdad'. Es un proceso de consulta de la comunidad donde se usan modelos para animar a los residentes a explorar y priorizar opciones para la acción (Gibson, 1998).

3.6. Monografías

Los estudios monográficos se usan para compilar datos descriptivos por medio del examen intensivo de un fenómeno en un individuo, grupo o situación. Los modelos monográficos se usan cuando se necesita una comprensión en profundidad muy detallada que sea holística, integral y contextualizada. Permiten establecer comparaciones entre diferentes sujetos dentro del mismo caso, entre casos y entre grupos de diferentes casos. Así, por ejemplo, en el contexto de una investigación basada en las escuelas, podría estudiarse cómo diferentes personas de la misma escuela tienen diferentes interpretaciones de una nueva iniciativa educativa, cómo diferentes escuelas la han llevado a la práctica o cómo contemplan la iniciativa los directores en contraste con los docentes o los alumnos. Las monografías son intensivas y, por tanto, pueden ser caras o complejas o exigir mucho tiempo, pero pueden aportar una comprensión muy profunda a la evaluación de políticas. Los datos observacionales y etnográficos pueden ser triangulados por otra gente que haya participado en la actividad observada o por otros investigadores (especialmente donde se han utilizado grabaciones de vídeo o audio).

Por tanto, los estudios monográficos son especialmente útiles a la hora de estudiar fenómenos infrecuentes o complejos. Robert Stake, entre otros, ha sido un firme partidario de las monografías y de que el evaluador elabore una 'descripción densa'. Piensa que los puntos de vista de los interesados son un elemento crucial de las evaluaciones y que los estudios monográficos son el mejor método tanto para representar las convicciones y valores de las personas interesadas como para presentar los resultados de una evaluación.

Stake arguye que existen realidades múltiples y que los puntos de vista de los interesados deben figurar en la evaluación, pero también sostiene que las personas interesadas no participan en la evaluación como querrían los teóricos de la participación. Se opone a la participación de las personas interesadas tal como se describe más arriba y argumenta que la evaluación es tarea del evaluador (Alkin, Hofsetter y Ai, 1998: 98), principalmente a través del trabajo monográfico, que puede incluir la observación, el análisis de documentos relativos a un caso particular, entrevistas en profundidad y otros métodos de recopilación de datos.

3.7. Observación de participantes

Una de las principales maneras que tiene la investigación social de entender una actividad, un grupo o un proceso es aproximarse lo más posible sin llegar a alterar su funcionamiento 'natural' (Hammersley y Atkinson, 1995). En un extremo, esto puede hacerse siendo un observador absolutamente desapegado de una situación social, trabajando de la manera más discreta posible, observando, escuchando y recordando detalles; en el otro extremo, uno puede unirse al grupo o actividad en cuestión y participar en él como miembro para aprender desde dentro. Esta opción puede o no implicar 'hacerse indígena', es decir, involucrarse

tanto en el grupo, la actividad o el proceso que se pierda la objetividad y la condición exógena. Evidentemente, entre estos dos extremos hay opciones: uno puede trabajar como observador-participante o como participante-observador, siendo la diferencia el grado de desapego e implicación que puede tener el investigador social.

3.8. Métodos participativos

Los métodos participativos hacen posible que las personas afectadas por un proyecto, un programa o una estrategia se involucren activamente en la toma de decisiones, y creen un vínculo con los resultados y recomendaciones de la evaluación. Son una herramienta útil para identificar los problemas durante los procesos de implementación y para aprender acerca de las condiciones locales y las perspectivas y prioridades de la población local, para así diseñar intervenciones más receptivas y sostenibles, aunque a veces se consideran métodos menos objetivos que los que dependen exclusivamente de evaluadores externos y pueden consumir demasiado tiempo de los agentes locales si es que se van a involucrar en profundidad en el proceso de evaluación. El punto de partida de la mayor parte del trabajo participativo y las valoraciones sociales es el análisis de la población afectada, que se utiliza para desarrollar una comprensión de las relaciones de poder, influencia e intereses de los diversos sujetos involucrados en una actividad, y para determinar quién debería participar en la evaluación y cuándo. Otros modelos comunes de los métodos participativos son la 'valoración de beneficiarios' (que implica el contacto sistemático con los beneficiarios del proyecto para identificar y diseñar iniciativas de desarrollo, localizar obstáculos para la participación y proporcionar información para mejorar los servicios y actividades) y el 'seguimiento y evaluación participativa' (en el que personas interesadas a diferentes niveles trabajan juntas para identificar problemas, compilar y analizar información y generar recomendaciones).

Al contrario que otros autores, como Scriven o Eisner, que consideran al evaluador un 'valorador', Guba y Lincoln (1989) consideran que las personas afectadas son quienes establecen primordialmente el valor. Este punto de vista se basa en la idea de que, lejos de haber una sola realidad, hay realidades múltiples basadas en las percepciones e interpretaciones de los individuos a quienes atañe el programa a evaluar. Así, Guba y Lincoln opinan que el papel del evaluador es facilitar las negociaciones entre individuos que reflejen esas realidades múltiples y abogan por los métodos participativos.

David Fetterman fue un paso más allá en *Empowerment Evaluation* (Fetterman *et al.*, 1996), donde describe la evaluación como un proceso que fomenta la autodeterminación entre los participantes en la evaluación del programa y que a menudo comprende "formación, facilitación, propugnación, iluminación y liberación". El objetivo de esta evaluación, que confiere poder a los participantes, es potenciar la autodeterminación en lugar de la dependencia, cosa que se consigue haciendo que, en lo esencial, los afectados por el programa efectúen sus propias evaluaciones. El evaluador externo sirve a menudo como consejero o auxiliar adicional, facilitando a los participantes los conocimientos y las herramientas necesarias para la evaluación continua y la atribución de responsabilidades. Para él, el punto final de la evaluación no es la valoración del programa; el valor y la valía no son estáticos: considera a la evaluación un proceso continuo que "puede desarrollarse para acomodar cambios en la población, los fines, las valoraciones y las fuerzas externas" (Fetterman, 1998).

4. ¿CÓMO SABER SI ALGO HA FUNCIONADO? DEFINIENDO LA BUENA EVALUACIÓN DE UNA POLÍTICA

La visión tradicional de la evaluación de políticas, surgida en los Estados Unidos en la década de 1960 –un periodo caracterizado por la necesidad urgente de evaluar los programas de la *Great Society* del gobierno estadounidense y por el optimismo acerca del conocimiento científico tras un periodo lleno de éxitos para las ciencias naturales–, se basa en la idea de que es posible dar respuestas científicas (objetivas) a las preguntas incluidas en cualquier proceso de evaluación.

En tiempos más recientes esta concepción ha sido atacada desde varios puntos, algunos de los cuales se han descrito ya en este artículo. El argumento básico de la mayoría de las críticas a la visión tradicional es que las presuposiciones bajo las cuales opera esta modelo de evolución no se sostienen. La evaluación es una empresa política y social, donde las diferencias de valor son relevantes, no sólo una tarea técnica y científica. Hay, además, múltiples perspectivas: esto no se debe sólo a la ambigüedad de los resultados de las políticas, programas e iniciativas en la práctica, sino también a las disensiones sobre qué tipo de criterios evaluativos son significativos o justos en una situación determinada (Majone, 1989). Subirats (1994) arguye que estas ambigüedades y estos problemas no pueden resolverse empleando simplemente más y mejores técnicas de medición. Los evaluadores 'pluralistas' argumentan que la evaluación debería utilizarse como un instrumento para crear confianza y consenso por medio de discusiones conjuntas iterativas, no para averiguar qué ha funcionado y qué no. Así, Subirats (1994) afirma que el evaluador "no debe actuar en solitario, decidiendo arbitrariamente si el programa del que se trata es bueno o malo; debe, más bien, actuar como mediador entre las diversas opiniones".

En cierto sentido, este argumento central es una proposición sorprendente. Si bien es cierto que pocos evaluadores sostendrían que las premisas de la evaluación tradicional son completamente válidas y menos aún descuidarían la importancia de tratar con los distintos afectados en el proceso de evaluación y escucharlos, parece excesivo sugerir que una evaluación realizada por un experto independiente será indefectiblemente 'arbitraria' si la acomete en solitario –si se respetan, varios criterios como la solidez y la fiabilidad del proceso de evaluación pueden controlar en gran medida la arbitrariedad– o que la función primordial del evaluador es mediar entre diferentes afectados más que elaborar una valoración de las políticas o los programas, de acuerdo con las condiciones contenidas en los términos de referencia del proyecto.

Esto se debe a varios motivos. En primer lugar, creer que la discusión sistemática puede conducir al consenso puede resultar excesivamente optimista en el contexto de muchas políticas. En segundo lugar, el marco temporal que requerirían las evaluaciones que verdaderamente aspirasen a este fin sería irrealizable, aun suponiendo que fuese posible. En tercer lugar, espera demasiado de los evaluadores. Existen, desde luego, otras instancias en que el diálogo entre diferentes afectados en el proceso de la política es más adecuado y se dispone de mayores recursos para llevar al consenso que durante un proceso de evaluación. Aunque las evaluaciones pueden –y deben– valorar la relevancia de las políticas, los programas y las iniciativas y de los debates que condujeron a la adopción de una política determinada (algo que a menudo pasan por alto los críticos pertenecientes a la tradición 'pluralista' de la evaluación[6]), no pueden absorber el proceso deliberativo de la creación de políticas, porque esa

función, habida cuenta de los rápidos cambios de dirección de las políticas en el mundo actual, llegaría de todas maneras demasiado tarde con excesiva frecuencia. En último lugar, pasa por alto que los evaluadores son contratados para llevar a cabo unas tareas especificadas en los términos de referencia de su proyecto de evaluación; pocas veces pueden permitirse el lujo de informar solamente sobre los diversos puntos de vista de los afectados y tratar de llegar a un consenso. En un contexto de disminución de los fondos públicos e incremento del número de agencias que compiten por estos fondos, los creadores de políticas y los demás responsables necesitan valoraciones de lo que ha funcionado y lo que no para informar y legitimar sus demandas y decisiones. Si la evaluación se acomete con rigor y los creadores de políticas la toman en serio (cosa que hacen cada vez más, al menos en el caso de programas e intervenciones de magnitud), el aspecto valorativo de la evaluación es de gran importancia. En él reside gran parte de la utilidad de la evaluación de políticas.

La elección entre distintos tipos de evaluación y distintos métodos de compilación de datos es, así pues, hasta cierto punto, una cuestión técnica, aunque condicionada por las limitaciones de los recursos y por los contextos políticos. Si bien, como se ha destacado más arriba, en este campo existe una gran variedad, hay varios criterios que pueden usarse para identificar las buenas evaluaciones de políticas que pueden ser de ayuda para los responsables de las decisiones. Las buenas evaluaciones se caracterizan por:

- Un conjunto definido de cuestiones de investigación que son lo bastante **específicas** como para poder ser llevadas a la práctica durante la investigación. Las cuestiones de investigación amplias o vagas conducen con facilidad a estudios insatisfactorios que, simplemente, no proporcionan nuevos conocimientos. Esto puede impedirse al principio del proceso evaluador dedicando más tiempo a definir qué hace falta saber. No obstante, es frecuente que durante el proceso de investigación aparezcan cuestiones adicionales, lo que significa que es posible que haya que revisar y enmendar las cuestiones iniciales a la vista de estos desarrollos.

- Ser **coherentes**. Debe haber coherencia entre las cuestiones de investigación y la población a estudiar; ésta debe ser la población que vaya a proporcionar la información más directa y profunda acerca del tema en cuestión. Como se ha comentado más arriba, todos los afectados deben participar en la evaluación, para mejorar su calidad y hacer sus resultados más interesantes para quienes pueden adoptar sus recomendaciones y producir un cambio.

- Ser **lógicas**. Debe existir una relación lógica entre las cuestiones de investigación y los métodos de compilación de datos utilizados –incluyendo el muestreo, sea aleatorio, deliberado o teórico– y una lógica subyacente a la distribución cronológica de los episodios de compilación de datos. Esto obliga a pensar cuidadosamente qué perspectiva (o perspectivas) acerca de lo que se está investigando puede resultar más reveladora.

- El uso de la **triangulación**. La triangulación consiste en combinar diversos tipos de datos –o, a veces, diversas maneras de observarlos– para responder a las cuestiones planteadas en la investigación. Denzin (1989) describe cuatro tipos de triangulación: la triangulación metodológica, que combina diferentes métodos de investigación; la triangulación de datos, que combina datos de más de una fuente; la triangulación de investigadores, que hace que más de un investigador observe los datos para replicar a las interpretaciones de otros investigadores; y la triangulación teórica, que contempla los datos desde diversas posturas teóricas para ver cómo de adecuadas resultan y comprender la manera en que la consideración de los datos desde diferentes presupuestos puede afectar a la manera en que se comprenden. De estos cuatro tipos, los dos primeros son los más usados en las evaluaciones gubernamentales.

- La atención a la **validez** interna y externa de los resultados. En *Experimental and Quasi-Experimental Designs for Research* (1966), Campbell y Stanley denominaron validez interna a la medida en la que un experimento se controla adecuadamente y *validez externa,* a la medida en que son ampliamente aplicables los resultados de un experimento. Por ejemplo, en los planes de evaluación, es a menudo imprescindible prestar atención al 'peso muerto' y las externalidades para obtener resultados válidos.

- La atención a la **fiabilidad** de los resultados, de manera que el instrumento de medición, sea cualitativo o cuantitativo, utilizado en la evaluación dé resultados coherentes, estables y uniformes durante sucesivas observaciones o mediciones efectuadas en idénticas condiciones.

- La atención al principio de **proporcionalidad,** que simplemente apunta a la necesidad de que el trabajo de evaluación corresponda a la escala de la intervención.

- La atención a la **objetividad** y la **integridad**. Los individuos que lleven a cabo el trabajo de evaluación deben carecer de impedimentos que empezcan su objetividad y han de actuar con integridad en sus relaciones con todas las personas afectadas.

- La producción de resultados **oportunos**, **relevantes** y **verosímiles**. Los resultados de la evaluación deben satisfacer las necesidades del organismo que la encargó y hacerse públicos en el momento más oportuno para contribuir a la toma de decisiones administrativas. La evidencia debe ser suficiente en relación con el contexto de la toma de decisiones; los resultados deben ser relevantes para los temas que se tratan y deducirse de la evidencia. Los resultados de la evaluación deben ser manifiestamente útiles para los gestores a la hora de mejorar el rendimiento e informar sobre los resultados obtenidos.

- El **estilo accesible**. Los informes deben ser concisos y estar claramente redactados; no deben incluir más información que la necesaria para una comprensión adecuada de los resultados, conclusiones y recomendaciones; deben presentar las conclusiones y las recomendaciones de manera que se deduzcan lógicamente de los resultados de la evaluación; y deben también exponer claramente los límites de la evaluación en cuanto a alcance, métodos y conclusiones.

Notas

1 Para una discusión más detallada de este tema, véase la última sección de este artículo.

2 Aunque las evaluaciones del impacto deberían cubrir los resultados duros y los blandos (más intangibles), la bibliografía al respecto tiende a centrarse en los primeros: una bibliografía reciente recopilada por Lloyd y O'Sullivan (2004) revelaba que, de hecho, hay muy pocas referencias a la medición de los resultados blandos o 'distancia recorrida'. Este hecho contrasta con los diversos modelos prácticos de medición de la distancia recorrida que se describen en su trabajo, lo cual indica que la bibliografía académica y de investigación de políticas acerca de este tema aún no está al mismo nivel que la práctica actual en la administración de proyectos.

3 Esta tarea no está exenta de problemas: por ejemplo, hay que considerar cuidadosamente qué programa 'alternativo' se ofrecerá al grupo de comparación (o control), dado que esto definirá la hipótesis. Idealmente, el grupo de control continuaría como lo habría hecho si no hubiese existido el programa. Con todo, hay motivos por los que esto puede no suceder:

1. El grupo de control puede enterarse de la existencia de la intervención, lo cual puede afectar a su comportamiento; por ejemplo, sus miembros pueden posponer sus actividades de búsqueda de empleo hasta reunir las condiciones para ser beneficiarios del nuevo programa.

2. Al grupo de control puede ofrecérsele los 'mejores' servicios disponibles en la actualidad como alternativa a la iniciativa de la política. Si, en ausencia del nuevo programa, no existen procedimientos para informar a la gente de estos servicios, el grupo de control no será comparable con la actual situación.

3. La intervención puede tener efectos que afecten indirectamente al grupo de control; por ejemplo, cambiando la actitud de los empleadores locales hacia los ex-delincuentes cuando se lleve a la práctica un programa de empleo para este grupo.

Otros problemas relacionados son el 'sesgo en la puesta en práctica' y el 'sesgo en la cola'. El sesgo en la puesta en práctica se da si, en el contexto de un ensayo, no se puede poner en práctica un programa como se pondría en un contexto más natural. El sesgo en la cola puede darse porque en un ensayo sólo un porcentaje de la población se ve afectado por la nueva intervención, y esto puede darle una ventaja injusta con respecto al grupo de control, cosa que no ocurriría si el programa se llevase a la práctica por completo.

Estos sesgos potenciales son muy difíciles, si no imposibles, de cuantificar. Los evaluadores tienen que esforzarse en minimizar los sesgos donde sea posible, pero idealmente también tendrían que ser capaces de emitir juicios informados acerca de cuán problemáticos pueden ser los sesgos.

4 Nuestras descripciones se basan en gran medida en Purdon (2002).

5 En gran parte de nuestra descripción seguimos a Davies (2004b).

6 Véase, por ejemplo, la afirmación de Subirats (1994: 9) según la cual "cuando se clasifica a una política como éxito o fracaso, esto a menudo indica que se la ha considerado desde una estrechez de miras centrada en la gestión, más preocupada por cumplir los fines internos de la política o por ejercitar un control administrativo efectivo que por la capacidad del programa para responder a las necesidades de los diversos individuos y grupos afectados". Esto pasa por alto que las evaluaciones de calidad –estén o no dentro del paradigma tradicional– también deben valorar la relevancia de las diversas iniciativas, así como su eficacia y su eficiencia. Esa valoración de la relevancia es en la actualidad un requisito para, por ejemplo, todas las evaluaciones que se acometen en la Comisión Europea.

REFERENCIAS BIBLIOGRÁFICAS

Adler, M. y Ziglio, E. 1996. *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health.* London: Jessica Kingsley Publishers.

Alkin, M. C., Hofstetter, C. H. y Ai, X. 1998. 'Stakeholder Concepts in Program Evaluation', en A. Reynolds y H. Walberg. *Advances in Educational Productivity, vol. 7.* Greenwich: JAI Press.

Blundell, R. y Costa Dias, M. 2000. 'Evaluation Methods for Non-Experimental Data', *Fiscal Studies,* 21 (4).

Boruch, R. F. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Newbury Park: Sage.

Bryson, A., Dorsett, R. y Purdon, S. 2002. *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies.* Working Papers, 4. London: UK Department for Work and Pensions.

Campbell, D. T. y Stanley, J. C. 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally College Publishing Company.

Cantrill, J. A., Sibbald, B. y Buetow, S. 1996. 'The Delphi and Nominal Group Techniques in Health Services Research', *The International Journal of Pharmacy Practice,* 4.

Chen, H. T. 1990. *Theory Driven Evaluations.* Thousand Oaks: Sage Publications.

Chen, H. T. y Rossi, P. H. 1983. 'Evaluating with Sense: A Theory-Driven Approach', *Evaluation Review,* 7 (3).

Connell, J. P. y Kubisch, A. C. 1995. 'Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects and Problems', en Fulbright-Anderson, K., Kubisch, A. C. y Connell, J. P. (eds.). *New Approaches to Evaluating Community Initiatives.* Washington: The Aspen Institute.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Light, R. J., Louis, T. A. y Mosteller, F. 1992. *Meta-Analysis for Explanation.* New York: Russell Sage.

Coote, A. y Lenaghan, J. 1997. *Citizens' Juries: Theory into Practice.* London: Institute for Public Policy Research.

Critcher, C. y Gladstone, B. 1998. 'Utilizing the Delphi Technique in Policy Discussion: A Case Study of a Privatized Utility in Britain', *Public Administration,* 76.

Davies, S., Elizabeth, S., Hanley, B., New, B. y Sang, B. 1998. *Ordinary Wisdom: Reflections on an Experiment in Citizenship and Health.* London: Kings Fund.

Davies, P. T. 2004a. 'Is Evidence-Based Government Possible?' Jerry Lee Lecture, pronunciada en el 4º Annual Campbell Collaboration Colloquium, Washington, 18-20 de febrero, 2004.

Davies, P. T. 2004b. *The Magenta Book. Guidance Notes for Policy Evaluation and Analysis.* London: UK Government Chief Social Researcher's Office.

Deeks, J. J., Altman, D. G. y Bradburn, M. J. 2001 'Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-Analysis', en Egger, M., Davey Smith, G. y Altman, D. G. (eds.). *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Publishing Group.

Denzin, N. K. 1989. *The Research Act: A Theoretical Introduction to Sociological Methods.* Englewood Cliffs: Prentice Hall.

Eisner, E. W. 1994 *The Educational Imagination: On the Design and Evaluation of School Programs.* New York, Toronto: Macmillan.

European Commission 1999. *Indicators for Monitoring and Evaluation: an Indicative Methodology.* The New Programming Period 2000-2006, Methodological Papers. Working Paper, 3. Brussels: DG XVI, Regional Policy and Cohesion, European Commission.

European Commission 2000. *The Mid-Term Evaluation of Structural Funds Interventions.* The New Programming Period 2000-2006, Methodological Papers. Working Paper, 8. Brussels: DG XVI, Regional Policy and Cohesion, European Commission.

Fetterman, D. S. 1998. *Ethnography: Step by Step.* Thousand Oaks: Sage.

Fetterman, D. M., Kaftarian, S. J. y Wandersman, A. 1996. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability.* Thousand Oaks: Sage.

Fishkin, J. 1995. *The Voice of the People. Yale:* Yale University Press.

Funnel, S. 1997. 'Program Logic: An Adaptable Tool for Designing and Evaluating Programs', *Evaluation News and Comment,* 6 (1).

Gibson, T. 1998. *The Do-ers' Guide to Planning for Real.* Neighbourhood Initiatives Foundation.

Greenberg, D. H. y Morris, S. 2003. *Large Scale Social Experimentation in Britain: What Can and Cannot Be Learnt from the Employment Retention and Advancement Demonstration.* Occasional Papers, 3. London: UK Government Chief Social Researcher's Office.

Greene, J. C., Benjamin. L. y Goodyear, L. 2001. 'The Merits of Mixing Methods in Evaluation', *Evaluation,* 7 (1).

Guba, E. G. y Lincoln, Y. S. 1989. *Fourth Generation Evaluation.* Newbury Park: Sage.

Hammersley, M. y Atkinson, P. 1995. *Ethnography: Principles in Practice.* London: Routledge.

Heckman, J. 1995 *Instrumental Variables: A Cautionary Tale.* Technical Working Paper, 185. Cambridge: NBER.

Heckman, J. J. y Smith, J. A. 1999. *The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies.* Working Papers, 6983. US National Bureau of Economic Research, Inc.

Lloyd, R. y O'Sullivan, F. 2004. *Measuring Soft Outcomes and Distance Travelled: A Practical Guide.* London: UK Department for Work and Pensions.

Majone, G. D. 1989. *Evidence, Argument and Persuasion.* New Haven: Yale University Press.

Marradi, A. 1990. 'Classification, Typology, Taxonomy', *Quantity and Quality,* XX (2).

Morton, S. 1999. *Systematic Reviews and Meta-Analysis.* Workshop Materials on Evidence-Based Health Care. San Diego, La Jolla: University of California.

Nagel, S. (ed.) 1990. *Policy Theory and Policy Evaluation: Concepts, Knowledge, Causes and Norms.* New York: Greenwood.

Owen, J. M. y Rogers, P. J. 1999. *Program Evaluation, Forms and Approaches.* London: Sage.

Park, A., Jowell, R. y McPherson, S. 1999. *The Future of the National Health Service: Results from a Deliberative Poll.* London: Kings Fund.

Patton, M. Q. 2002. *Qualitative Research & Evaluation Methods.* London: Sage

Pawson, R. y Tilley, N. 1997. *Realistic Evaluation.* London: Sage.

Purdon, S. 2002. *Estimating the Impact of Labour Market Programmes.* Working Paper, 3. London: UK Department for Work and Pensions.

Rodgers, P. J., Petrosino, A., Huebner, T. A. y Hacsi, T. A. 2000. *New Directions for Evaluation: Program Theory in Evaluation-Challenges and Opportunities.* San Francisco: Jossey Bass Publishers.

Rosenbaum, P. R. y Rubin, D. B. 1983. 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika,* 70.

Rossi, P. H., Freeman, H. E. y Lipsey, M. W. 1999. *Evaluation: A Systematic Approach.* Newberry Park: Sage Publications.

Scriven, M. 1972. 'Pros and Cons about Goal-Free Evaluation', *Evaluation Comment, 3.*

Seargeant, J. y Steele, J. 1998. *Consulting the Public: Guidelines and Good Practice.* London: Policy Studies Institute.

Slavin, R. E. 1984. 'Meta-Analysis in Education: how has it been used?', *Educational Researcher,* 13.

Slavin, R. E. 1986. 'Best Evidence Synthesis: An Alternative to Meta-Analysis and Traditional Reviews, *Educational Researcher,* 15.

Subirats, J. 1994. *Policy Instruments, Public Deliberation and Evaluation Processes.* Estudio-Working Paper, 1994-51. Madrid: Instituto Juan March de Estudios Avanzados en Ciencias Sociales.

Treasury Board of Canada Secretariat 2001. 'Evaluation Policy'. Disponible en http://www.tbs-sct.gc.ca/pubs_pol/dcgpubs/TBM_161/ep-pe1_e.asp#_Toc505657347.

Weiss, C. H. 1997. 'Theory-Based Evaluation: Past, Present and Future', *New Directions for Evaluation,* 76.

White, H. 2005. Challenges in Evaluating Development Effectiveness. IDS Working Papers, 242. Brighton: University of Sussex.

White, C., Elam, G. y Lewis, J. 1999. *Citizens' Juries: an Appraisal of their Role.* London: Cabinet Office.

World Bank 2004. *Monitoring and Evaluation: Some Tools, Methods and Approaches.* Washington: World Bank Evaluation Operations Department.