DOCUMENTO DE TRABAJO

Guía metodológica de evaluación de programas educativos

La redacción de este informe estuvo a cargo de María Pía Pirelli.

Comisión Directiva del INEEd: Alex Mazzei (presidenta), Pablo Cayota, Alejandro Maiche, Limber Elbio Santos, Marcelo Ubal y Oscar N. Ventura.

Dirección Ejecutiva del INEEd: Mariano Palamidessi.

Montevideo 2016.

Edificio Los Naranjos, planta alta, Parque Tecnológico del LATU.
Av. Italia 6201, Montevideo, Uruguay.
(+598) 2604 4649 – 2604 8590.
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: INEEd (2016), *Guía metodológica de evaluación de programas educativos*, INEEd, Montevideo.

En la elaboración de este material se ha buscado que el lenguaje no invisibilice ni discrimine a las mujeres y, a la vez, que el uso reiterado de /o, /a, los, las, etcétera, no dificulte la lectura.

Agradecimientos: El INEEd agradece especialmente los aportes y comentarios de Cecilia Llambí.

Introducción

Este documento constituye un insumo para planificar la evaluación de un programa educativo. Para ello presenta de forma breve, referenciando las fuentes y exponiendo ejemplos relativos al campo de estudio, las distintas modalidades en que un programa puede ser evaluado.

En primer lugar, se considera como una etapa necesaria de trabajo el considerar la propia "evaluabilidad" del programa. Luego se presentan diferentes abordajes y metodologías que se corresponden con diversas metodologías para evaluar distintos aspectos de los programas. Se explican distintos tipos de evaluaciones: teoría del cambio, cadena de resultados y evaluación basada en teoría; evaluación del diseño; evaluación de la implementación; evaluación de impacto; evaluación del costo-efectividad y, por último, un panorama de otras formas de clasificar a las evaluaciones. El documento cierra con un apartado sobre los usos de los resultados de la evaluación de programas.

Evaluación de la evaluabilidad de un programa

La evaluación de la evaluabilidad es un proceso sistemático para describir la estructura de un programa y para analizar la plausibilidad y factibilidad del logro de los objetivos y su adecuación para la evaluación en profundidad (Smith, 2005: 136).

La EA (evaluability assesment) incluye un mínimo de precondiciones que tienen que cumplirse antes de proceder a cualquier esfuerzo por evaluar un programa o una política (Rossi, Freeman y Lipsey, 1999: 157). La EA generalmente incluye tres actividades primarias: 1) descripción del "modelo del programa" con una especial atención a sus objetivos y metas, 2) valoración de cuán bien definido y evaluable es el modelo del programa, y 3) identificación de los intereses sobre la evaluación de las partes involucradas y los probables usos de los hallazgos (Rossi, Freeman y Lipsey, 1999: 157).

Los evaluadores en esta etapa comienzan a trabajar a partir de la concepción sobre el programa que aparece en los documentos oficiales, pero buscan ir más allá, para llegar a una visión más certera de cómo el programa es en realidad, así como entender las cuestiones del programa que realmente le importan a las partes interesadas (Rossi, Freeman y Lipsey, 1999: 157). Muchas veces a partir de este diagnóstico inicial ya son recomendados a los responsables de la política cambios al programa en cuestión. La EA puede revelar fallas en el sistema de entrega de los servicios, en la definición de la población objetivo o la necesidad de reconceptualizar el programa (Rossi, Freeman y Lipsey, 1999: 157).

Por otra parte, para la Organización para la Cooperación y el Desarrollo Económicos (OCDE) la evaluabilidad es el grado en que una actividad o proyecto puede ser evaluado de forma fiable y creíble (Davies, 2013: 7). Sin embargo, el concepto es utilizado en dos sentidos diferentes y complementarios. Uno es el de la evaluabilidad

"en principio", que mira la naturaleza del diseño del proyecto, incluyendo su teoría del cambio y le pregunta si es posible evaluarlo como está descrito. El otro es el de la evaluabilidad "en la práctica" y mira a la disponibilidad de datos pertinentes, así como los sistemas y capacidades que hacen que los datos estén disponibles (Davies, 2013: 7). Es necesario examinar la utilidad probable de una evaluación. Los resultados de una evaluación de la evaluabilidad deberían tener consecuencias para el diseño de una evaluación, para el diseño de un marco de evaluación y monitoreo, o para el diseño del proyecto en sí (Davies, 2013: 8).

Al final de la EA se tendría que poder dar cuenta de tres cosas. La primera es la medida en que la teoría de cómo el programa se espera que funcione se alinea con cómo es implementado y percibido en el campo. En segundo lugar, de la plausibilidad de que el programa vaya a dar resultados positivos tal como fue concebido y puesto en práctica actualmente. Por último, de la viabilidad y los mejores enfoques para su evaluación adicional.

Teoría del cambio, cadena de resultados y evaluación basada en teoría

Las políticas o programas públicos buscan modificar algún aspecto de la realidad. En primer lugar, se identifica un aspecto de la realidad social que se desea modificar como respuesta a una necesidad o demanda de cambio, ya que se valora otro escenario como mejor. En segundo lugar, los decisores de políticas tienen una teoría (a veces explícita, otras veces implícita) sobre cómo ciertas acciones podrían modificar esa realidad en el sentido deseado (teoría del cambio). Por último, se crea una intervención pública (a través de una ley, un programa u otro medio) que parece resultar la mejor opción posible para realizar ese cambio (Castro y Pirelli, 2014: 1).

Una teoría del cambio “es una descripción de cómo se supone que una intervención conseguirá los resultados deseados. Describe la lógica causal de cómo y por qué un proyecto, un programa o una política lograrán los resultados deseados o previstos” (Gertler y otros, 2011: 22).

A partir de esta teoría se construyen las hipótesis del programa. Su explicitación es fundamental para armar el plan de acción, para la evaluación de diseño y para la de impacto. De fondo está la idea de causalidad: qué secuencia de eventos o actividades y bajo qué condiciones y supuestos se daría el cambio, lo que permite generar una intervención lógica.

Una teoría de la política expresa la “promesa” de causalidad entre los recursos, los instrumentos y aquellos objetivos específicos. Por definición, cualquier teoría de la política da una versión simplificada de la realidad, y a la vez dirige la percepción, la interpretación y la evaluación (Van der Knaap, 2004: 18).

Las teorías políticas guían las percepciones, los pensamientos y las acciones. Ellas dan foco pero también pueden distorsionar la visión del mundo. La teoría es entendida, en un sentido más amplio, como la forma en que las personas organizan su conocimiento sobre el mundo y, por lo tanto, el mundo mismo. El conocimiento

esquemático y los supuestos incorporados en las teorías sirven como base para nuestro entendimiento del mundo, incluyendo los programas de política pública (Van der Knaap, 2004: 18).

La función de una teoría mientras se desarrolla la política es la de focalizar y reducir la complejidad, lo que permite al responsable centrarse en los aspectos más importantes y en la forma de cómo lograrlo. El expresar la teoría de la política le da al responsable de ella una posición clara en los debates y provee un marco de referencia para todas las partes interesadas. Luego de la etapa de debate, la teoría se constituye como el punto de partida de la acción colectiva sobre la cual se pueden hacer los ajustes provisionales por medio del monitoreo y puede tener lugar el “aprendizaje para la mejora” o el “aprendizaje para la innovación” (Van der Knaap, 2004: 24).

Es así que existe el enfoque de las evaluaciones basadas en teorías (*theory-based evaluation*), el cual puede ser definido como el análisis y la valoración de la contribución de las estrategias de intervención para resolver o controlar un problema específico. El punto de partida tradicional de la evaluación basada en la teoría es proporcionada por los objetivos y las hipótesis en que se basa un programa de política determinada.

Si la teoría no es explícita, siempre está la posibilidad de que exista una teoría oculta detrás de una serie de medidas o instrumentos de medición. La tarea para el evaluador es entonces tratar de reconstruir las teorías "innatas" que guiaron la toma de decisiones y la acción (Van der Knaap, 2004: 26).

Una teoría de un programa sería buena si representa el conocimiento necesario para que aquel obtenga los resultados deseados, o sería una teoría pobre cuando no se obtienen los resultados esperados aun implementando de forma correcta el programa (Rossi, Freeman y Lipsey, 1999: 155). Por lo tanto, uno de los aspectos a evaluar cuando se está realizando la evaluación de un programa es valorar cuán buena es la teoría que tiene por detrás, en particular qué tan bien está formulada y si ella presenta una forma plausible y un plan factible para mejorar las condiciones de su población objetivo. Sin embargo, para poder evaluar la teoría es necesario que esta haya sido expresada de forma clara (Rossi, Freeman y Lipsey, 1999: 155).

Todo programa encarna una concepción de las estructuras, funciones y procedimientos apropiados para lograr los objetivos. Esta concepción constituye la lógica o plan del programa, que se llama *teoría del programa*. Esta explica por qué el programa hace lo que hace y provee la base lógica para esperar que realizando determinado camino de acciones se logren los resultados deseados (Rossi, Freeman y Lipsey, 1999: 156).

La evaluación basada en la teoría, generalmente, en primera instancia gira en torno al análisis y evaluación de la medida en que los programas de la política han dado lugar a sus objetivos originales. Una segunda característica de este tipo de evaluación es una mejor comprensión de los mecanismos causales subyacentes. En este caso, la pregunta central es: ¿fueron los supuestos en que el programa de política se basó "correctos" o no? (Leeuw, 1983 en Van der Knaap, 2004: 17).

La evaluación basada en teoría permite determinar el diseño de proyectos complejos, y mejorar la planificación y la gestión (Banco Mundial, 2004: 10). Tiene varias semejanzas con el concepto de marco lógico, pero permite una comprensión más profunda del funcionamiento de un programa o actividad: la "teoría del programa" o "lógica del programa". Pero no se supone una relación sencilla y lineal entre causa y efecto.

Un ejemplo en educación

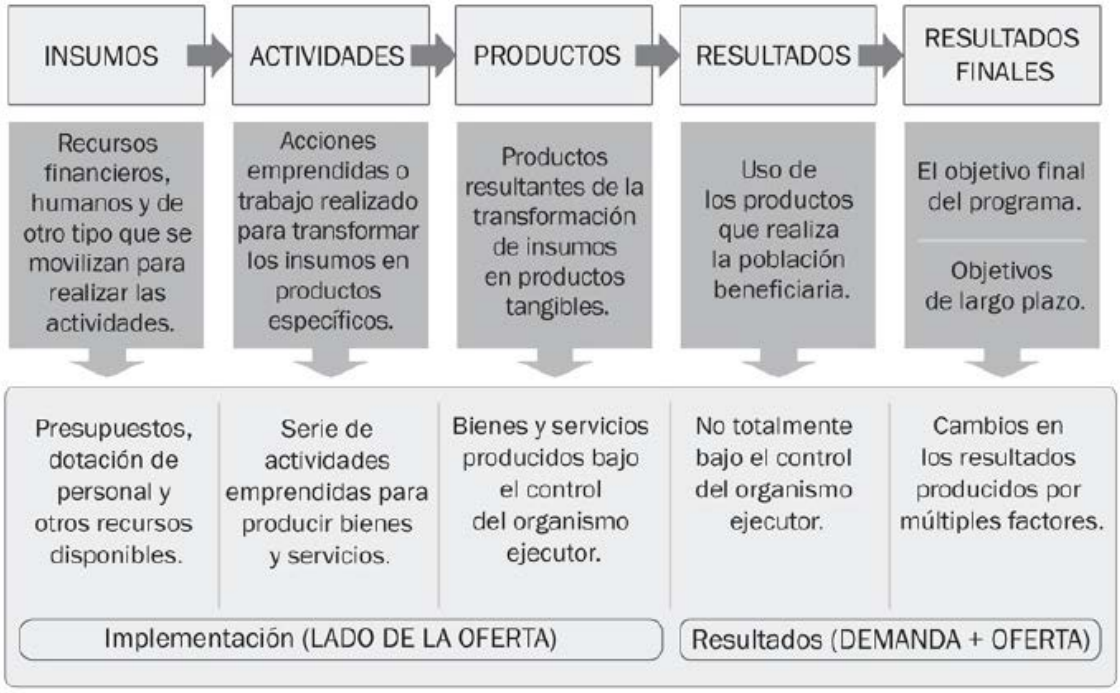
> "Por ejemplo, el éxito de un programa gubernamental para mejorar la alfabetización incrementando el número de maestros podría depender de numerosos factores, como la disponibilidad de aulas y libros de texto, las probables reacciones de los padres, directores de escuela y alumnos, la preparación y moral del personal docente, los distritos en que se deben ubicar los maestros adicionales, la fiabilidad del financiamiento público, etc. Si se consigue establecer cuáles son los factores determinantes del éxito y la forma en que se interrelacionan, se puede decidir qué pasos deben supervisarse a medida que avanza el programa, para ver si encuentran confirmación en los hechos. Ello permite determinar cuáles son los factores definitivos del éxito. Y cuando los datos revelan que estos factores no se han conseguido, una conclusión razonable es que el programa tiene pocas probabilidades de conseguir sus objetivos".

Fuente: Banco Mundial (2004: 10).

La teoría del cambio, que señala la cadena causal de la política, puede plasmarse de diversas formas tales como modelos teóricos, modelos lógicos, marcos lógicos, modelos de resultados y cadenas de resultados. Todas ayudan a discernir entre los logros del programa y las condiciones o influencias ajenas al programa (Gertler y otros, 2011: 24).

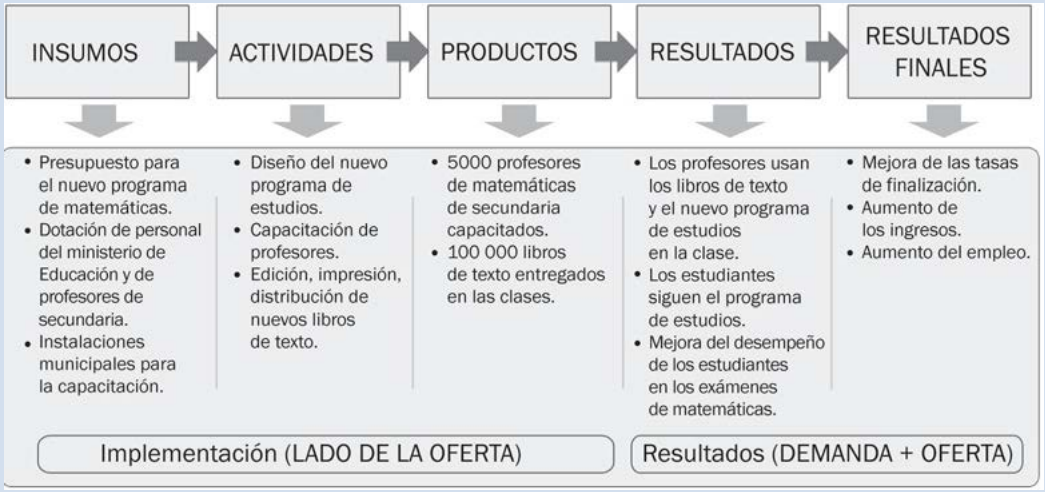
Una cadena de resultados establece la secuencia de insumos, actividades y productos de los que se espera que mejoren los resultados, y los resultados finales. Incluye una definición lógica y plausible de cómo esto sucede (Gertler y otros, 2011: 24). Tiene tres partes principales: implementación, resultados y suposiciones, y riesgos. La implementación refiere a los aspectos que puede supervisar directamente el organismo ejecutor para medir el desempeño del programa. Los resultados, por su parte, no dependen solo del control directo del proyecto, sino también de cambios de comportamiento de los beneficiarios del programa. En otras palabras, dependen de las interacciones entre el lado de la oferta (implementación) y el lado de la demanda (beneficiarios). Estos son los aspectos que se someten a la evaluación de impacto para medir la efectividad. Por otra parte, las suposiciones y riesgos "hacen referencia a los riesgos que pueden afectar a la consecución de los resultados esperados y a cualquier estrategia para solventar dichos riesgos. Incluyen cualquier evidencia en la bibliografía especializada sobre la lógica de causalidad propuesta y las suposiciones en que se basa" (Gertler y otros, 2011: 24)

Una cadena de resultados da una definición lógica y plausible de cómo una secuencia de insumos, actividades y productos relacionados directamente con el proyecto interactúan y establecen las vías por las que se logran los impactos.



Fuente: Gertler y otros (2011: 25, gráfico 2.1).

Un ejemplo en educación



Fuente: Gertler y otros (2011: 26, gráfico 2.2).

Evaluación del diseño

Un mayor énfasis en el diseño de políticas ayuda a garantizar que las actuaciones previstas representan un medio realista y viable para lograr los objetivos de la política (Hallsworth, Parker y Rutter, 2011: 6). Al evaluar el diseño del programa o política se quiere analizar si están claramente identificados el problema social que busca atacar, la justificación del tipo de intervención y las herramientas a utilizar (señalar qué antecedentes o bajo qué teorías se espera que reviertan los procesos señalados como

problemáticos), los objetivos y metas que se proponen, y si este conjunto de elementos guarda una coherencia lógica entre sí.

Como se trata del punto de partida de las políticas o programas y el análisis de la cadena causal que hay detrás de ellos, el papel de la teoría es central. Si no existe una coherencia de punta a punta en el diseño, no se puede suponer que tendrá los resultados esperados.

Los puntos que deben ser analizados necesariamente son: la especificación del problema o necesidad que el programa busca resolver, mencionando las principales causas y su respaldo empírico; cuán apropiada es la estrategia de abordaje; la consistencia en la lógica causal vertical entre los objetivos esperados y los bienes y servicios que presta el programa; la definición y estimación de la población potencial y objetivo del programa; la asignación de recursos y presupuesto; y el diseño de la estructura organizacional.

Un ejemplo en educación

A continuación se ilustran algunos puntos del análisis del diseño a partir de un programa educativo.

- Identificación del problema. Un programa educativo "X" es creado frente a la necesidad de elevar el nivel de capital humano promedio de una sociedad, como demanda del sector empresarial frente a la escasez de mano de obra calificada. Se justifica en base a la creencia de que por encontrarse la economía en un ciclo expansivo, aumentando la tasa de empleo, muchos jóvenes prefieren abandonar sus estudios para ingresar tempranamente al mercado laboral.

- Estrategia planteada. Este programa se apoya en dos herramientas para lidiar con esta realidad: becas económicas y tutorías para la finalización de estudios de educación media. Todas orientadas a la población con extraedad.

- Supuestos. Esto tiene como supuestos que: la culminación de la educación media provee de herramientas necesarias para mejorar la calidad del capital humano o que habilita de forma exitosa estudios terciarios posteriores, y el factor económico es determinante para la revinculación o permanencia de los estudiantes en el sistema educativo. En una evaluación de diseño convendría comparar el monto de la beca con el costo de oportunidad de estos jóvenes (por ejemplo, los salarios obtenidos por los jóvenes de esta edad con este nivel educativo en el mercado laboral).

- Definición de la población potencial. Esta es definida particularmente para cada uno de los productos que se ofrecen y los requisitos a la entrada del programa. La población objetivo en este caso se ve limitada por el presupuesto asignado al programa cada año.

Evaluación de la implementación u operacional

Hay dos razones principales por las que los programas fallan: porque falla la teoría en la que se basa o porque falla la implementación (es decir que nunca llega a implementarse adecuada o suficientemente como para poder medir su efectividad) (Patton, 2008).

La evaluación de la implementación se focaliza en determinar si el programa tiene todas sus partes, si las partes son funcionales al programa y si el programa está operando como se supone que debería hacerlo. En un inicio permite clarificar "qué es el programa" en la realidad, para poder luego avanzar hacia la pregunta más compleja de si realmente funciona (Patton, 2008: 308 y 309).

En este nivel se busca evaluar el nivel de implementación y los procesos del programa, es decir, una vez puestas en marcha las actividades señaladas en el diseño del programa. Este tipo de evaluación busca responder preguntas del tipo: ¿cuáles son las características claves del programa?, ¿quiénes participan?, ¿la organización cuenta con un equipo de trabajo bueno y capacitado?, ¿están las responsabilidades bien asignadas?, ¿qué hace el personal?, ¿están siendo completadas las tareas de los intermediarios a tiempo?, ¿a qué porcentaje de la población objetivo se está llegando y de qué forma?, ¿qué experimentan los participantes?, ¿qué está funcionando y qué no? (JPAL: 5; Patton, 2008)

La evaluación operacional analiza cuán efectivamente están siendo implementados los programas y si hay brechas entre lo planificado y lo realizado. Generalmente refiere a una evaluación retrospectiva sobre la base de los objetivos iniciales del proyecto, los indicadores y los objetivos del marco de monitoreo y evaluación (Khandker, Koolwal y Hussain, 2010: 16). Este tipo de evaluación permite identificar aprendizajes para nuevos diseños de políticas. Al ubicarse al nivel de la ejecución y tener el detalle de las operaciones del programa se puede determinar qué es lo que funciona del programa y evitar confusiones.

Uso de la información
• Se utiliza para la toma de decisiones de los responsables del programa/política. • Provee información que permite asignar y redistribuir recursos a la población objetivo. • Permite interpretar los hallazgos de una evaluación de impacto o resultados y determinar si las fallas se dan en la teoría o en el proceso de implementación.

Por otra parte, la evaluación de implementación permitiría identificar y comparar las discrepancias entre el diseño y la implementación real, pero también las discrepancias entre distintas unidades de implementación (como aulas, escuelas, localidades). Una de las cuestiones, entonces, es pensar la fidelidad versus la adaptación. Siempre que se disemina un programa es muy difícil mantener la fidelidad al diseño. Cuánta

adaptación y qué debe ser adaptado son buenas preguntas de evaluación (Patton, 2008).

Esto se ve claro cuando se quiere copiar una intervención que fue catalogada como "buena práctica". El programa puede apoyarse en evaluaciones formativas tanto si persigue el objetivo de mantenerse fiel al diseño original como si prefiere realizar una adaptación exitosa de este a entornos locales.

¿Qué factores interactúan en la implementación? Algunos ejemplos son las adaptaciones de los implementadores en el terreno dados los desafíos que emergen en la práctica, los intereses políticos y de las burocracias. Sucede también que el protocolo de trabajo no es del todo preciso cuando se inicia el programa, dando lugar a diversas interpretaciones y acciones de los actores que se encuentran implementando para poder subsanar las incógnitas. Ello da lugar a distintas modalidades de implementación o niveles de fidelidad de la implementación que deben ser tomadas en cuenta para evaluar los resultados.

La implementación entonces puede ser leída como una adaptación incremental al contexto local y sus complejidades emergentes. Por tanto, se puede decidir evaluar la fidelidad o decidir evaluar el nivel y tipo de adaptación local. Esto tiene consecuencias para el diseño metodológico de la evaluación.

Es difícil pensar hoy en un estudio de evaluación que no incluya algún aspecto de evaluación de implementación, ya que es útil a varios propósitos. Mejora la rendición de cuentas del programa documentando las actividades y esfuerzos, provee evidencia objetiva de que el programa está siendo entregado como estaba previsto y permite a los responsables tomar decisiones informadas acerca del diseño del programa y dirección de la política (Love, 2004: 96, en Patton, 2008: 323).

El punto clave es que sin información sobre la operativa actual del programa y de los mecanismos causales, los tomadores de decisiones se encuentran limitados en interpretar los datos del desempeño para poder realizar mejoras o juicios (Patton, 2008: 331).

Existen distintas evaluaciones de implementación en relación a los objetivos que persiguen con respecto al uso que se les quiere dar; pero todos estos enfoques pueden ser utilizados.

Tipo de evaluación de implementación
1) **Evaluación de esfuerzos**. La evaluación de esfuerzos, insumos y acceso se centra en la documentación de la cantidad y la calidad de las actividades que tienen lugar y los recursos disponibles para las actividades del programa, así como en la determinación de si la población objetivo verdaderamente está recibiendo el servicio.
2) **Monitoreo de programas.** La rutina de información de la gestión es un proceso típicamente interno a los programas, muchas veces apoyado en un software específico, y en contraste con ciertos parámetros establecidos que permitan determinar el progreso de la implementación.
3) **Evaluación de procesos.** Esta evaluación se focaliza en las dinámicas internas y las operaciones actuales de un programa en un intento de entender las fortalezas y debilidades. La mira está en cómo se produce un resultado más que el resultado en sí mismo. Una evaluación de proceso puede brindar una retroalimentación muy útil en la etapa de desarrollo del programa, así como posteriormente en la difusión y diseminación de un programa efectivo.
4) **Evaluación de componentes.** Implica una valoración formal de cada parte del programa, es decir, cada esfuerzo operacional separado, así como los vínculos entre ellas. Lo que cambia es la unidad de análisis y esta es una de sus principales ventajas, ya que aumenta las posibilidades de generalización de los resultados y de comparación con otros programas.
5) **Especificaciones del tratamiento**. Supone identificar y medir precisamente qué hay en un programa que se supone tiene un efecto. Las especificaciones del tratamiento revelan los supuestos causales que apuntalan las actividades del programa, por tanto, está intrínsecamente vinculada con la teoría. Especificar el tratamiento implica ir más allá de la etiqueta o título que lleve el programa.

Fuente: traducción propia de Patton (2008).

Impacto

Las evaluaciones de impacto buscan identificar si los cambios que se han producido luego de la implementación de un programa son atribuibles a este, así como poder cuantificar la magnitud de ese efecto, descartando la posibilidad de que factores distintos al programa en estudio expliquen el impacto observado. A partir de la información generada se puede decidir acerca de eliminar las intervenciones menos eficaces o escalar las que muestren buenos resultados:

> La pregunta básica de la evaluación de impacto es esencialmente un problema de inferencia causal. Evaluar el impacto de un programa sobre una serie de resultados es equivalente a evaluar el efecto causal del programa sobre ellos […] Las evaluaciones de impacto nos ayudan a atribuir causalidad al establecer empíricamente en qué medida cierto programa, y solo ese programa, ha contribuido a cambiar un resultado (Gertler y otros, 2011: 33 y-34).

La dificultad de este tipo de evaluaciones radica en que hay que comparar la situación de la población beneficiaria después de la intervención con la situación de la misma población si no hubiese existido o participado en la intervención: el contrafactual. El contrafactual refiere a cuál habría sido el resultado para un participante en el programa en ausencia del programa.

El desafío clave en una evaluación de impacto es el encontrar un grupo de personas que no participaron, pero que son lo suficientemente parecidas como para medir "cómo estarían los participantes si no hubiesen recibido el programa" (JPAL: 6). Es decir, grupos de comparación, también llamados grupos de control. Identificar grupos de comparación válidos es el punto central de la evaluación de impacto; sin una estimación válida del contrafactual es imposible conocer el impacto de un programa (Gertler y otros, 2011: 34).

Deben cumplirse al menos tres condiciones para que el grupo de comparación sea válido con respecto al grupo de tratamiento: ambos grupos deben ser idénticos en ausencia del programa, los grupos deben reaccionar de la misma manera al programa (no debe existir sesgo de selección) y ambos grupos no pueden estar expuestos de manera diferente a otras intervenciones durante el período de la evaluación (Gertler y otros, 2011: 38).

En lo que respecta a las formas de estimar el impacto, estas son de dos tipos: la de intención de tratar y la de tratamiento en tratados. La de intención de tatar refiere a cuando esta se calcula con las unidades a las que se ha ofrecido el programa, independientemente de si han participado o no en él. Esto es particularmente importante en los casos que se quiere conocer el impacto promedio sobre la población en la cual se ha focalizado el programa. Por otra parte, la estimación del impacto del tratamiento en tratados solo toma en consideración a los beneficiarios a quienes se les ofreció el programa y que efectivamente han participado de él (Gertler y otros, 2011: 39).

Por último, las dos siguientes estimaciones de contrafactual son falsas: 1) comparar *antes y después* los resultados de los mismos participantes del programa y 2) comparar *con y sin (tratamiento)* entre unidades que deciden inscribirse y unidades que no lo hacen (Gertler y otros, 2011: 40).

Hay varios métodos para realizar una evaluación de impacto y cada método viene acompañado de sus propios supuestos. Las principales técnicas son: el experimento aleatorio, el experimento natural, el método de emparejamiento, el método de variables instrumentales y el método de regresión discontinua. En la siguiente tabla se esbozan algunas características de los métodos.

Métodos de evaluación de impacto y algunas de sus características	
Experimento aleatorio	
La asignación aleatoria de un programa a los beneficiario entre una población elegible numerosa permite hacer una estimación robusta del contrafactual, lo que es clave para la evaluación de impacto. Además, la asignación aleatoria también es una manera justa y transparente de asignar los recursos escasos de un programa. Con este método es confiable interpretar el verdadero impacto del programa como la diferencia entre el resultado bajo tratamiento (el resultado medio del grupo de tratamiento asignado aleatoriamente) y la estimación del contrafactual (el resultado medio del grupo de comparación asignado aleatoriamente). Hay dos situaciones donde es viable aplicar este método de evaluación de impacto: cuando la población elegible es superior al número de plazas disponibles en el programa y cuando es necesario ampliar un programa gradualmente hasta que cubra a toda la población elegible. La asignación aleatoria puede hacerse a nivel de individuos, hogares, comunidades o regiones, lo cual depende de dónde y cómo se implementa el programa (Gertler y otros, 2011: 50, 51, 53, 55, 56 y 60).	
No experimentales	
Diferencias en diferencias	El método de diferencias en diferencias consiste en aplicar una doble diferencia, mediante la comparación de los cambios a lo largo del tiempo (antes y después) en la variable de interés entre una población inscrita en un programa (el grupo de tratamiento) y una población no inscrita (el grupo de comparación). Este método permite tener en cuenta cualquier diferencia constante en el tiempo entre ambos grupos. La utilización conjunta de dos falsos contrafactuales (comparaciones antes-después y comparaciones inscritos-no inscritos) permite generar una mejor estimación del contrafactual. Tiene la ventaja de poder aplicarse cuando las reglas de asignación del programa sean menos claras y no requiere que los grupos cuenten necesariamente con las mismas condiciones previas a la intervención (Gertler y otros, 2011: 95 y 96). El contrafactual que se estima aquí es el cambio en los resultados del grupo de tratamiento. Para aplicar diferencias en diferencias solo hace falta medir los resultados del grupo que recibe el programa (el grupo de tratamiento) y del grupo que no lo recibe (el grupo de comparación) antes y después del programa (Gertler y otros, 2011: 95, 96).
Método de emparejamiento (*matching*)	El emparejamiento utiliza “grandes series de datos y técnicas estadísticas complejas para construir el mejor grupo artificial de comparación posible para el grupo de tratamiento”. Para dicha construcción se basa en las características observadas de los individuos en la línea de base, suponiendo que no existen diferencias en las variables inobservadas entre los grupos que estén asociadas también con los resultados de interés (Gertler y otros, 2011: 107, 110). “Si la lista de características relevantes observadas es demasiado grande, o si cada una de ellas asume múltiples valores, puede ser difícil identificar una pareja para cada una de las unidades del grupo de tratamiento. Conforme aumenta el número de características o dimensiones en función de las cuales quiere emparejar las unidades inscritas en el programa, puede toparse con lo que se denomina ‘la maldición de las

	dimensiones'" (Gertler y otros, 2011: 107). Para subsanar este problema se puede aplicar el método de emparejamiento de las propensiones a participar (*propensity score matching*), donde para "cada unidad del grupo de tratamiento y del conjunto de no inscritos, se computa la probabilidad o propensión de que participa en el programa mediante los valores observados de sus características, la denominada 'puntuación de la propensión'" (Gertler y otros, 2011: 108).
Variables instrumentales	Este método se utiliza cuando presumiblemente existe sesgo de selección de los participantes de un programa debido a características inobservables (autoselección al tratamiento). Para ello necesita cumplirse el supuesto de que exista una variable, llamada instrumental, que está correlacionada con la participación en el programa, pero que no es una determinante directa de la variable de resultado. Hay que evidenciar que esta variable seleccionada es relevante y exógena. El estimador de variables instrumentales identifica un efecto local y no un efecto promedio, lo cual tiene limitaciones en términos del alcance de la interpretación del estimador (Bernal y Peña, 2011).
Regresión discontinua	Se utiliza cuando los programas sociales utilizan un índice continuo de elegibilidad con una puntuación límite claramente definida para determinar quién tiene derecho a participar y quién no. Bajo estas condiciones se puede utilizar el diseño de regresión discontinua (DRD) para medir el impacto (Gertler y otros, 2011: 82). "El diseño de regresión discontinua estima los impactos medios locales en torno al umbral de elegibilidad en el punto en el que las unidades de tratamiento y de comparación son más similares. Conforme el umbral está más cerca, las unidades a ambos lados se asemejan más. De hecho, sumamente cerca de la puntuación límite, las unidades a ambos lados del límite serán tan parecidas que la comparación será tan buena como si se hubieran elegido los grupos de tratamiento y de comparación mediante asignación aleatoria" (Gertler y otros, 2011: 91).

Fuente: elaboración propia a partir de Gertler y otros (2011).

Un ejemplo en educación

¿Pueden las campañas informativas generar una mayor conciencia y participación local en la educación primaria en India?

Desafío para la política pública. Pese al aumento en la matriculación de las escuelas primarias en el mundo, la calidad de la educación sigue siendo baja en muchos países. Para revertir esto, una de las respuestas ha sido implicar más a las comunidades locales en la supervisión y participación, a pesar de que existe poca evidencia rigurosa de que esto funcione o de cómo puede ser alentada esta participación. ¿Es más efectiva la acción directa de las comunidades para enseñar a los niños a leer?

Contexto de la evaluación. En 2001 el gobierno estableció Comités de Educación (*Village Education Committees* o VEC) en todos los pueblos. Los VEC están conformados por la cabeza del gobierno electo de la villa (*pradhan*), el director de la escuela local y tres padres nominados por la comunidad. Estos comités son responsables de monitorear el desempeño de la escuela, reclamar los fondos públicos y contratar otro profesor con contrato en caso de exceso de estudiantes. Sin embargo, en 2005 una encuesta reveló que los integrantes de los VEC no tenían claros los mecanismos de funcionamiento ni las atribuciones que tenían.

Detalles de la intervención. Se realizó un trabajo conjunto entre Pratham (una ONG local) y el Banco Mundial, para lo cual los investigadores diseñaron tres intervenciones que se asignaron aleatoriamente a 280 pueblos. Contrastaron esto con la acción directa para mejorar el aprendizaje, externa a los canales oficiales.

- Intervención 1. En 65 pueblos, los trabajadores en terreno mantuvieron por dos días una serie de conversaciones consultivas sobre el estado de la educación y sobre el conocimiento existente acerca del VEC, en pequeños grupos en toda la comunidad. Finalizadas estas, se realizó una reunión con la comunidad, en la que se instó a las personas a solicitar información del VEC, proporcionada por Pratham. El personal también entregó a los miembros del VEC un folleto con información acerca de sus funciones y responsabilidades.
- Intervención 2. Además de todos los pasos detallados anteriormente, las comunidades de otros 65 pueblos recibieron formación y motivación para realizar pruebas, con el objetivo de observar si los niños podían leer textos simples y solucionar problemas aritméticos básicos. Los voluntarios elaboraron una "tarjeta de informe" para cada comunidad, que se presentó en la reunión con la comunidad.
- Intervención 3. Además de los dos pasos anteriores, los funcionarios de Pratham enseñaron a los voluntarios de otros 65 pueblos una técnica simple para ayudar a que los niños aprendan a leer. Se instó a los voluntarios a dictar clases de lectura después de la escuela y el personal regresó en promedio siete veces para realizar formación en el trabajo. El objetivo era utilizar materiales diseñados de Pratham y voluntarios locales para complementar el plan de estudio normal y mejorar la alfabetización entre los niños del pueblo.
- 85 pueblos no recibieron tratamiento, los que sirvieron como comparación.

Resultados y lecciones de política pública.

- Impacto en los vacíos de la información. El efecto promedio de los tres tratamientos fue un aumento de 7,8% en los miembros de VEC que sabían que tenían acceso a fondos públicos y un aumento de 13% de miembros que recibieron una formación adecuada. Además, la probabilidad de que los padres conocieran la existencia de un VEC en su comunidad aumentó un 2,6%.
- Impacto en el compromiso. A pesar de estas mejoras en la generación de conciencia, existió una pequeña diferencia entre el desempeño del VEC de los pueblos en

tratamiento y de comparación. Además, la intervención no aumentó el nivel de compromiso de los padres con las escuelas.

- Impacto en la lectura. El niño promedio en un pueblo que recibió la Intervención 3, que era completamente analfabeto al inicio, tuvo 7,9% más probabilidades de leer al menos letras. Quienes solo podían leer letras al inicio fueron 3,5% más propensos a leer al menos párrafos o palabras y 3,3% más propensos a leer historias si se encontraban en un pueblo que recibió la Intervención 3. Si asumimos que todas las mejoras en los pueblos fueron consecuencia de las clases de lectura, entonces se deben haber observado grandes mejorías en la lectura de los niños que asistieron a las clases. Particularmente, los niños que no podían leer en el punto de partida y asistieron a las clases finalmente pudieron leer letras y el 98% de quienes podían leer palabras o párrafos pudo leer historias.

  Esta fue la única intervención que realmente mejoró los resultados educacionales, al potenciar a las personas para mejorar la enseñanza en sus propias comunidades. Esto sugiere que permitir la acción local, la cual no depende de la participación de un grupo grande, puede ser un medio para afectar directamente los resultados educativos.

Fuente: Banerjee (2010). Adaptación del resumen de la investigación, disponible en la web de JPAL https://www.povertyactionlab.org/es/evaluation/campanas-informativas-educacion-primaria-India

Costo-efectividad

A todo responsable de política pública debiera interesarle poder escoger entre programas que muestren iguales resultados, el que tenga menores costos, ya que eso podría permitir utilizar los fondos excedentes para atender a más beneficiarios o utilizarlos en otras áreas de interés.

Un análisis costo-beneficio cuantifica los beneficios y costos de una actividad y los pone en la misma medida métrica (a menudo en una unidad monetaria). Se trata de responder la pregunta: ¿está el programa produciendo suficientes beneficios para compensar los costos? O en otras palabras, ¿la sociedad será más rica o más pobre después de realizar esta inversión? Este enfoque es más útil cuando hay múltiples tipos de beneficios y se ha acordado monetizarlos (JPAL: 7).

En términos generales, un análisis de costo-efectividad directo toma el impacto de un programa y lo divide por el costo del programa. Sin embargo, este tipo de análisis puede complejizarse al considerar, por ejemplo, los costos directos e indirectos, los eventuales ahorros de costos si el programa redunda en ello, los beneficios directos y las externalidades.

Un ejemplo en educación

Comparación de estrategias para aumentar la asistencia escolar en Kenia

Mediante la evaluación de una serie de programas en circunstancias similares, se puede comparar el costo-efectividad de diferentes estrategias para mejorar indicadores de resultados, como la asistencia escolar. En Kenia, la organización no gubernamental International Child Support Africa implementó una serie de intervenciones educativas que incluyen el tratamiento contra parásitos intestinales, la provisión gratuita de uniformes escolares y la oferta de desayunos escolares. Cada una de las intervenciones se sometió a una evaluación de impacto basada en una asignación aleatoria y a un análisis de costo-beneficio. La comparación entre ambos estudios generó observaciones interesantes sobre cómo aumentar la asistencia escolar.

Un programa que suministró tratamiento contra los parásitos intestinales a niños en edad escolar aumentó la asistencia alrededor de 0,14 años para cada niño tratado, con un costo estimado de US$ 0,49 por niño. Esto equivale a unos US$ 3,50 por cada año adicional de asistencia a la escuela, a los que hay que agregar las externalidades experimentadas por los niños y los adultos que no asisten a la escuela en las comunidades que se beneficiaron de la reducción de la transmisión de los parásitos.

Una segunda intervención, el Programa de Patrocinio de Niños, redujo el costo de la asistencia a la escuela mediante el suministro de uniformes escolares a los alumnos de siete centros seleccionados aleatoriamente. Las tasas de abandono escolar disminuyeron drásticamente en las escuelas tratadas y, después de cinco años, se estimó que el programa había aumentado los años de escolarización un promedio del 17%. Sin embargo, incluso con supuestos más optimistas, el costo del aumento de la asistencia a la escuela mediante el programa de uniformes escolares se estimó en alrededor de US$ 99 por cada año adicional de asistencia.

Finalmente, un programa que suministró desayunos gratuitos a niños de 25 centros de preescolar seleccionados aleatoriamente generó un aumento del 30% en la asistencia a las escuelas del grupo de tratamiento, con un costo estimado de US$ 36 por cada año adicional de escolarización. Los resultados de los exámenes también mejoraron con desviaciones estándar de alrededor de 0,4 puntos, siempre que el profesor hubiera estado bien formado antes del programa.

Aunque intervenciones similares pueden tener distintos objetivos en términos de resultados, como los efectos de la desparasitación para la salud o el logro escolar, además del aumento de la participación, la comparación de una serie de evaluaciones realizadas en el mismo contexto puede determinar qué programas lograron el objetivo deseado con el menor costo.

Fuente: Gertler y otros (2011: 12).

Otra clasificación de las evaluaciones

Diferencias entre evaluación y monitoreo

Es importante diferenciar el monitoreo de la evaluación, términos que muchas veces son usados indistintamente. Por monitoreo se entiende el seguimiento sistemático de las acciones del programa y sus productos o resultados, a efectos de hacer ajustes para mejorar la gestión y dar cuentas en forma pública de lo realizado. Implica el relevamiento de información y su reporte en forma continua y sistemática. El punto de referencia para el monitoreo es el plan de acción establecido para el programa bajo el supuesto de que este es el mejor camino para alcanzar los objetivos buscados (Mokate, 2000: 3).

Una evaluación, en cambio, involucra la construcción de juicios valorativos más amplios sobre el programa, a partir de la información recogida, orientados a la toma de decisiones. La evaluación se puede entender como un proceso para determinar el mérito o el valor de algo, por lo que involucra la identificación de criterios de referencia relevantes que proporcionen un elemento de comparación para el análisis y la construcción de juicios de valor. La evaluación se extiende más allá del monitoreo porque reconoce que el plan de acción de un programa es una hipótesis con respecto al camino hacia el logro de los objetivos (Mokate, 2000: 3 y 4).

Evaluaciones *ex ante* y *ex post*

Las evaluaciones *ex ante* (antes del hecho) predicen el impacto de un programa utilizando datos antes de su implementación, como, por ejemplo, las microsimulaciones. Por su parte, las evaluaciones *ex post* (después del hecho) analizan los resultados que arroja el programa luego de su implementación. Clásicamente dentro de este último tipo entran las evaluaciones que comparan a la población beneficiaria antes y después de la intervención, o las que comparan participantes versus no participantes (Khandker, Koolwal y Hussain, 2010: 7).

Evaluaciones cuantitativas y cualitativas

Si bien muchas veces las evaluaciones pueden buscar cuantificar resultados y medir el impacto de las políticas, para lo cual utilizan datos recogidos a través de encuestas o pruebas, la información cualitativa, como la comprensión del contexto sociocultural, de las instituciones locales, de los programas y de los detalles de los participantes es esencial para una evaluación cuantitativa con sentido (Khandker, Koolwal y Hussain, 2010: 18 y 19). "La información cualitativa sirve, entre otras cosas, para anticipar la heterogeneidad, para intuir algunas de las características y atributos de los individuos que afectan o confunden los impactos y, en general, para interpretar mejor los resultados de la evaluación de impacto cuantitativa" (Bernal y Peña, 2011: 6). Una mezcla de métodos cualitativos y cuantitativos (un enfoque de métodos mixtos) puede ser útil para obtener una visión global de la eficacia del programa.

Uso de la evaluación

Desde el inicio es importante que el evaluador pueda, en diálogo con el responsable del programa o política, conocer las expectativas y poder determinar qué tipo de evaluación es la más útil para determinada circunstancia. A esto se le suma el trabajo que es necesario realizar con la información proveniente de los hallazgos de la evaluación. Esto puede leerse, siguiendo el planteo de Patton (2008), en torno a cuatro procesos: el análisis, la interpretación, los juicios y las recomendaciones.

1. El análisis implica organizar los datos crudos para que tengan un formato entendible que muestre los patrones básicos de los resultados. Este constituye los hallazgos empíricos de la evaluación. En este nivel hay que ser muy claro con las definiciones acerca de qué se está midiendo para que no haya lugar a malos entendidos. También es importante que el evaluador, a pesar de haber utilizado técnicas sofisticadas para llegar a los resultados del estudio, sea capaz de, con creatividad, comunicar de forma sencilla y entendible los hallazgos. Por ello es más recomendable que en su presentación haga énfasis en los aspectos destacados, y las técnicas sofisticadas sean explicadas en un apéndice o pie de página. Por tanto, hay que distinguir la complejidad del análisis de la claridad de la presentación.

Por otra parte, en general todo tipo de análisis de evaluación termina siendo en algún sentido comparativo. El punto es seleccionar una base de comparación adecuada que no aparezca como arbitraria o artificial. Por lo tanto, la comparación para que sea apropiada y relevante debe ser decidida entre el evaluador y las partes interesadas en el programa.

Los resultados de un programa pueden ser comparados con:
1. Los resultados de un programa “similar” seleccionado
2. Los resultados del mismo programa en el año anterior u otro periodo de tiempo
3. Los resultados de una muestra aleatoria o representativa de programas en el terreno
4. Los resultados de algún programa especial de interés
5. Los objetivos fijados del programa
6. Los propios objetivos de los participantes
7. Estándares externos de deseabilidad
8. Estándares de mínima aceptabilidad
9. El ideal de rendimiento del programa
10. Supuestos hechos por el equipo u otros responsables

Fuente: Patton (2008: 485).

2. La interpretación implica determinar qué significan los hallazgos, cuán significativos son y qué explica los resultados, yendo más allá de los datos, agregando el contexto. Para ello, es necesario tener en cuenta cuatro dimensiones a trabajar con los actores del programa:

- los números y la información cualitativa debe ser interpretada para que signifique algo y se pueda ver cómo debería ser aplicada;
- los datos son indicadores o representaciones imperfectas de lo que es el mundo;

- los datos estadísticos y los cualitativos contienen errores, los investigadores ofrecen probabilidades, no absolutos; y
- buscar la significancia que salta a la vista (Patton, 2008: 487).

3. Los juicios brindan los valores que permiten determinar el mérito y si los resultados son positivos o negativos, buenos o malos, deseables o no. Los hallazgos en sí mismos no permiten determinar nada si no se poseen valores o estándares con respecto a los que comparar. Estos criterios hay que intentar que sean primeramente establecidos por los usuarios de la evaluación, generalmente los responsables del programa, política o los financiadores. Aunque podría establecerlos el evaluador, es mejor que este facilite el trabajo de discusión de los estándares con los actores del programa para generar un mayor compromiso y apropiación de la evaluación.

4. Las recomendaciones implican determinar acciones a partir de los hallazgos, por lo que necesariamente están basadas en la evidencia. Generalmente este punto es la parte más visible de las evaluaciones. Por ello Patton (2008) realiza una serie de diez sugerencias sobre estas:

- el foco de las recomendaciones debe ser negociado y clarificado con los interesados y los financiadores de la evaluación como parte del diseño;
- las recomendaciones deben provenir de forma clara y fundamentada en los hallazgos de la evaluación;
- se debe distinguir entre distintos tipos de recomendaciones, es importante ponderarlas para que no quede una lista muy extensa que no permita resaltar las de mayor peso;
- algunos tomadores de decisiones prefieren recibir múltiples opciones más que recomendaciones que solo defiendan un único curso de acción;
- se deben discutir los costos, beneficios y desafíos de implementar las recomendaciones;
- es necesario focalizarse en acciones controlables por los usuarios previstos;
- hay que trabajar con los actores para analizar las implicaciones políticas de las recomendaciones;
- se debe ser cauto en la redacción de la evaluación, recomendaciones importantes pueden perderse por el uso de un lenguaje vago y obtuso (hay que evitar palabras que sean confusas y distraigan acerca del mensaje central);
- es importante dejar tiempo para hacer un buen trabajo acerca de las recomendaciones, tiempo para desarrollarlas de forma colaborativa con los interesados, y tiempo para pilotear recomendaciones acerca de su claridad, comprensibilidad, factibilidad, y significatividad;
- es preciso desarrollar estrategias para que las recomendaciones sean tomadas seriamente (hay que ayudar a los responsables a tomar las decisiones, para ello puede ser útil realizar sesiones de trabajo donde se incluya una asignación de responsabilidades para el seguimiento de las acciones y un calendario para las implementaciones).

Bibliografía

AGEV-OPP (2011) Evaluaciones de Diseño, Implementación y Desempeño Guía Metodológica y Notas técnicas, Montevideo.

Banco Mundial (2004), *Seguimiento y evaluación: instrumentos, métodos y enfoques*, Departamento de Evaluación de Operaciones, Washington D.C.

Banerjee, Abhijit (2010), "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India" en American Economic Journal: Economic Policy 2(1): 1-30.

Berk, Richard y Rossi, Peter (1998), *Thinking about Program Evaluation*, Sage, Thousand Oaks.

Bernal, Raquel y Peña, Ximena (2011), *Guía práctica para la evaluación de impacto*, Universidad de los Andes, Facultad de Economía, Bogotá.

Castro, Lucía y Pirelli, María Pía (2014), "La importancia de evaluar los programas educativos", en Boletín institucional del INEEd, INEEd, Montevideo.

Davies, Rick (2013), *Planning Evaluability Assessments. A Synthesis of the Literature with Recommendations*, Department for International Development, Working Paper 40, Cambridge.

Gertler, Paul; Martínez, Sebastián; Premand, Patrick; Rawlings, Laura B. y Vermeersch, Christel (2011), *La evaluación de impacto en la práctica*, Banco Mundial, Washington D.C.

Hallsworth, Michael; Parker, Simon y Rutter, Jill (2011), *Policy Making in the Real World. Evidence and Analysis*. Institute for government, Londres.

JPAL (s/f) *Introducción a las evaluaciones*, disponible en https://www.povertyactionlab.org/sites/default/files/documents/introduccion-evaluaciones.pdf.

Khandker, Shahidur R.; Koolwal, Gayatri B. y Hussain, A. Samad (2010), *Handbook on Impact Evaluation: Quantitative Methods and Practices*, Banco Mundial, Washington D.C.

Merino, Marisa (2007), *La evaluabilidad: de instrumento de gestión a herramienta estratégica en la evaluación de políticas públicas*, Papeles de Evaluación nº 7, Agencia de Evaluación y Calidad, Ministerio de Administraciones Públicas de España.

Mokate, Karen (2000), "El monitoreo y la evaluación: herramientas indispensables de la gerencia social", en *Diseño y gerencia de políticas y programas sociales*, BID-INDES.

Patton, Michael Quinn (2008), *Utilization-Focused Evaluation*, SAGE, Thousand Oaks.

Rossi, Peter; Freeman, Howard y Lipsey, Mark (1999), *Evaluation. A Systematic Approach*, Sage, Thousand Oaks.

Smith, M. F. (2005), "Evaluability Assessment", en S. Mathison (ed.), *Encyclopedia of evaluation*, Sage, Thousand Oaks.

Van der Knaap, Peter (2004), "Theory-Based Evaluation and Learning: Possibilities and Challenges", en *Evaluation*, SAGE, Thousand Oaks.