Ignacio Moral-Arce*

EVALUACIÓN *EX POST* DE UNA INTERVENCIÓN PÚBLICA

Este trabajo analiza la evaluación ex post de una intervención pública, analizando los productos logrados y recursos empleados, y posteriormente una valoración de los resultados logrados. La evaluación de impacto es un tipo de análisis que aísla el efecto del programa sobre la variable de resultado. A continuación se muestran las diferentes preguntas de evaluación que se pueden realizar y las metodologías más habituales. Finalmente se presentan las evaluaciones de contrafactuales, donde el sesgo de selección es un elemento de gran importancia, lo que afecta a la utilización de diseños experimentales o cuasi experimentales.

Ex-post evaluation of a public intervention

This paper analyses the ex-post evaluation of a public intervention, analysing the outputs achieved and resources used, and subsequently an assessment of the outcomes achieved. Impact evaluation is a type of analysis that isolates the effect of the programme on the outcome variable. The various evaluation questions that can be posed and the most common methodologies used are shown below. Lastly, counterfactual evaluations are presented where selection bias is a hugely important factor, which influences the use of experimental or quasi-experimental designs.

Palabras clave: *impacto, contrafactual, sesgo selección, dif in dif, emparejamiento.*
Keywords: *impact, counterfactual, selection bias, dif in dif, pairing.*
JEL: *C54, C9, H43.*

1. Introducción

Cuando se diseña una intervención pública[1], el objetivo último buscado es solucionar, o mitigar en la medida de lo posible, un problema detectado en la sociedad, por lo que, cuando la política está orientada a obtener resultados, una vez que esta ha sido ejecutada, las evaluaciones que se presentan en este tema pretenden determinar si el objetivo inicial se logró. Si bien la evaluación de los programas no es algo nuevo, las limitaciones de recursos y las demandas políticas para una mayor rendición de cuentas y transparencia han puesto mayor énfasis sobre las políticas basadas en la evidencia, que enfatiza que el diseño y éxito de cualquier política debe basarse en hechos y ser capaz de demostrar y, cuando sea posible, medir los «resultados», el «valor por dinero» y la «efectividad». Precisamente, las evaluaciones *ex post*, centradas en resultados e impactos, investigan los cambios que provoca una intervención basada en la evidencia, tratando de determinar si un programa es eficaz,

* Dirección de Estudios. Instituto de Estudios Fiscales.
Versión de 22 de febrero de 2019.
DOI: https://doi.org/10.32796/ice.2019.907.6783

[1] A lo largo del texto, con el término de «intervención pública» se quiere denominar a cualquier iniciativa llevada a cabo por el sector público —en sus diferentes niveles— que pretenda generar un efecto en la sociedad.

y particularmente la evaluación de impacto (EI) busca demostrar que los resultados esperados se derivan de las actividades del programa.

Este artículo se estructura del siguiente modo. En el siguiente apartado se presentan las características de la evaluación *ex post*. A continuación se analiza la evaluación de recursos y productos empleados en el programa estudiado, para pasar en el apartado 4 a la evaluación de resultados. La evaluación de impacto se analiza en el apartado 5 y las diferentes metodologías en el apartado 6. Se cierra el trabajo mostrando las evaluaciones de impacto con contrafactuales y los diseños más habituales, y el apartado de conclusiones.

2. Características de la evaluación *ex post*

La evaluación *ex post* tiene como objetivo principal valorar el nivel de eficacia y eficiencia logrado por la intervención pública. Hay que destacar dos características que presenta este tipo de evaluación. Por un lado, teniendo en cuenta la cadena de valor de un programa público dado por el Esquema 1. Este tipo de evaluaciones se centra en analizar tres fases de la cadena de valor: recursos, productos y resultados. Por otra parte, el momento de tiempo en que se realiza, produciéndose una vez que la intervención ha finalizado.

Los productos son los bienes y servicios generados por las actividades de un programa, como las clases impartidas, empresas ayudadas, o los participantes de un curso. La finalidad de cualquier producto que se genera mediante una actuación debe ser producir los resultados deseados para los participantes del programa, por lo que tienen que ser actuaciones cuantificables. Los resultados son los beneficios para los participantes después de su participación en un programa: la última etapa en la cadena de valor. Este resultado previsto se define como la dimensión específica del bienestar y el progreso de los individuos, y su cambio es el que precisamente motiva la intervención realizada.

La evaluación *ex post* puede ser de producto o resultado e impacto (Moral-Arce, Paniagua, Rodríguez y Rodríguez, 2016). La distinción fundamental entre ambas reside en su finalidad. La primera persigue mejorar la eficacia y eficiencia operacional del proyecto, mientras que la segunda se centra en la efectividad lograda por la intervención mediante la determinación de los cambios que este ha producido en la variable de resultado. Por lo tanto, se pueden diferenciar dos tipos de categorías dependiendo de la fase en la que se centra el estudio: *i*) análisis focalizado en productos y servicios distribuidos. En este caso se puede realizar un análisis de eficacia[2], teniendo en cuenta la cobertura y focalización, y/o un análisis de eficiencia; y *ii*) análisis centrado en resultados, donde se puede diferenciar entre una evaluación de resultados propiamente dicha, y/o una evaluación de impacto.

3. Evaluación de recursos y productos: eficacia y eficiencia

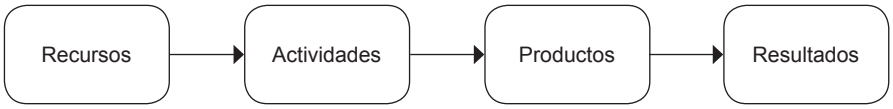
Una vez finalizada una intervención pública, es necesario determinar el grado de ejecución, estudiando qué se ha obtenido y sus características más importantes, empleando indicadores de producto y recursos. Este tipo de evaluaciones a menudo las usan los administradores y gestores como puntos de referencia para medir el éxito de una intervención pública (Kirkpatrick, 1995), y ayuda a contestar diferentes tipos de preguntas. En concreto, se deben tener en cuenta los siguientes aspectos:

— Eficacia: se analiza el cumplimiento de todo lo establecido en los objetivos del programa. Se debe uno cerciorar de que todas las actuaciones se han llevado a cabo y, tras ejecutarlas, que su cumplimiento se transforme en la consecución de los objetivos del programa. Se centra en analizar los recursos disponibles,

[2] «eficacia» y «efectividad» son términos que se emplean para indicar el grado de consecución de los objetivos establecidos por la organización, ya sean estos de productos o de resultados. Para evitar confusión sobre en qué fase de la cadena de valor se está analizando, en este artículo se entenderá que la evaluación de eficacia se centra en productos. Por lo tanto, analiza el nivel de cumplimiento en la cantidad de bienes y servicios que se habían programado inicialmente, mientras que la evaluación de efectividad se centra en resultados.

ESQUEMA 1

CADENA DE VALOR



FUENTE: Elaboración propia.

productos generados, a cuánta población se ha llegado, el tiempo que se ha consumido, etc.

— Eficiencia: tras evaluar la eficacia del programa es importante determinar cómo se ha producido la consecución de los objetivos, es decir, establecer la relación entre los recursos utilizados por el programa y los productos (bienes y/o servicios). Se analizan fundamentalmente los costes medios efectivos, así como la comparación con los costes planificados en la fase de diseño de la intervención.

— Calidad: se valora la forma en que se han entregado dichos bienes, comprobando si la política se presta de una forma adecuada y satisfactoria para la población.

4. Evaluación de resultados: efectividad

El objetivo último buscado por cualquier intervención pública es solucionar un problema detectado en la sociedad, por lo que este enfoque se caracteriza por medir el éxito de una intervención pública en términos de los resultados logrados. La primera aproximación consistirá en comparar las metas de resultados establecidas en la fase inicial de la planificación/programación respecto a los valores realmente logrados (ejecutadas), para identificar la existencia de posibles desviaciones. Sin embargo, se recomienda ser muy prudente a la hora de utilizar la información generada en esta evaluación de resultado, ya que el hecho de que un indicador evolucione en sentido contrario al programado originalmente no indica, ni mucho menos, que la política no esté funcionando, aunque es posible. Sencillamente, puede ser que la situación macroeconómica sea tan mala que no sea posible mejorar. Por lo tanto, si no se logran los resultados esperados, pueden existir varias razones, y solo en algunos casos se debe a que el proyecto no ha funcionado correctamente.

El problema de la atribución

Los programas efectivos son aquellos que marcan la diferencia en el cumplimiento de los objetivos esperados en los resultados de la intervención pública. Al tratar de medir el rendimiento de un programa, nos enfrentamos a dos problemas. Con frecuencia se puede, no sin cierta dificultad, medir si estos resultados están ocurriendo realmente o no. La pregunta más difícil generalmente es determinar qué contribución hizo el programa específico en cuestión al resultado. ¿Cuánto del éxito (o el fracaso) en el resultado se puede atribuir al programa? ¿Cuál ha sido el aporte del programa? ¿Qué influencia ha tenido? A pesar de la dificultad de la medición, la atribución es un problema que no se puede ignorar al tratar de evaluar el desempeño de los programas gubernamentales. Sin una respuesta a esta pregunta, poco se puede decir sobre el valor del programa, ni se puede proporcionar asesoramiento sobre direcciones futuras. Quizás incluso sin el programa, los cambios observados en los resultados habrían sucedido, o habrían ocurrido en un nivel inferior o posterior. En la mayoría de los casos, hay otros factores en juego además del impacto de las actividades del programa, como otras acciones o programas

gubernamentales, factores económicos o tendencias sociales que pueden tener un efecto en los resultados. La pregunta es: ¿cómo podemos demostrar que un programa está generando una diferencia? En este análisis de «atribución» se quiere determinar la parte del valor del resultado que se debe exclusivamente al programa analizado. Por ese motivo surgen las «evaluaciones de impacto», cuyo objetivo fundamental es precisamente aislar el efecto que tiene el recibir la intervención pública del resto de factores existentes que pueden afectar a la variable de interés analizada.

5. La evaluación de impacto

La EI analiza si una intervención pública genera los impactos buscados en la población de beneficiarios. Este enfoque se puede simplificar estudiando qué hubiera sucedido con y sin intervención pública, lo que permite aislar el efecto del programa, es decir, su atribución sobre el resultado del programa: lo que exclusivamente se debe al programa y no a otros elementos (Imas y Rist, 2009).

La atribución causal es un elemento esencial de cualquier EI, y permite que una evaluación informe no solo que se produjo un cambio, sino también que se debió, al menos en parte, al programa o política que se está analizando. Por tanto, esta aproximación implica que no se puede afirmar que un cambio es un impacto a menos que haya un vínculo demostrado entre él y la intervención.

Preguntas de evaluación de impacto

Una EI debe concentrarse en una pequeña cantidad de preguntas clave específicas de evaluación. Dependiendo del objetivo de la evaluación, existen diferentes tipos de preguntas de «impacto», o incluso más probablemente una combinación de estas. Algunas evaluaciones buscan respuestas precisas a preguntas muy claras; mientras que otras querrán entender si un programa ha tenido algún tipo de efecto; o estarán más interesadas en los motivos o explicaciones de por qué ha sucedido eso (Stern, 2015). Las preguntas típicas que hacen las evaluaciones de impacto son:

¿En qué medida se puede atribuir un impacto específico a la intervención?

Estudia cuánto del resultado observado se debe a la intervención. Esta pregunta está contestando «si/no» hay impacto, y «cuánto» es el efecto. Para contestar a esta pregunta se emplea el enfoque clásico del contrafactual, por lo que se debe disponer de un grupo de control, porque este tipo de análisis requiere de la comparación entre la existencia del programa (grupo de tratamiento) y sin él (grupo de control). Estas aproximaciones permiten demostrar que se está produciendo algún tipo de conexión causal.

¿La intervención pública ha tenido efecto (ha supuesto una diferencia)?

Cada vez más, en la actualidad, esta segunda pregunta genera más interés en los responsables de las políticas. Esto se debe a que la gran mayoría de intervenciones públicas solo son parte de otros factores, incluidas otras políticas públicas con similares objetivos realizadas por otros agentes, además de factores externos, que tendrán diferentes grados de influencia en los resultados. Por lo tanto, reconocer la contribución de otros es más realista que buscar evidencia de atribución exclusiva. Esto también es consistente con los desarrollos metodológicos en las ciencias sociales que se centran en la causalidad múltiple, los «paquetes causales» y las «causas contributivas». Estos desarrollos se basan en entender qué «causas» pueden ser necesarias pero no suficientes para conducir a un cambio, e incluso puede ser que exista más de una manera de lograr un objetivo similar.

¿Cómo ha logrado la intervención tener efecto?

Esta pregunta se centra en el «cómo» y «por qué» se han logrado esos efectos. Trata de contestar cuál

fue el proceso por el que la intervención condujo o contribuyó a los resultados. El objetivo es aprender para mejorar el éxito, por lo que se necesitan cierto tipo de explicaciones y análisis, y en estas circunstancias el conocimiento que se tiene sobre la intervención pública se vuelve importante, y pretende abrir la «caja negra» que conecta causas y efectos.

¿Esta intervención pública podría funcionar en otro entorno?

Esta pregunta también responde a «por qué». Pretende determinar qué otros factores necesitan estar presentes a lo largo de la intervención para generar los resultados observados, o qué otros factores son necesarios/suficientes para que la intervención funcione. Existen situaciones en las que resulta evidente que programas similares no siempre conducen al mismo resultado en todos los lugares. Esta es la razón por la que la «teoría» es un elemento fundamental para juzgar el éxito y el fracaso de las políticas. En EI, como en la investigación científica, la explicación depende en última instancia de buenas teorías, recurriendo a teorías sociales, económicas y comunitarias más amplias para interpretar datos complejos y, a menudo, confusos o incluso contradictorios.

6. Metodológicas de evaluación de impacto

En este apartado se presentan algunas de las aproximaciones que permiten contestar a las preguntas que se han indicado previamente. Estas incluyen: *i*) evaluación basada en la teoría; *ii*) evaluación realista; *iii*) análisis comparativo cualitativo; *iv*) análisis de contribución; y *v*) evaluación contrafactual, entre otros. Pasamos a continuación a describir cada uno de ellos (Befani, 2016)[3].

[3] Dentro de las evaluaciones más cualitativas, merece la pena destacar los trabajos realizados por la extinta AEVAL, actualmente Instituto para la Evaluación de Políticas Públicas, y también los del Instituto de Evaluación en Cataluña de IVALUA.

Evaluación basada en la teoría

La evaluación basada en la teoría es un enfoque en el que se presta atención a las teorías que se emplean en la formulación de políticas, realizadas por los gestores de programas u otras partes interesadas, lo que implica la recopilación de supuestos e hipótesis, empíricamente verificables, que estén vinculados entre sí.

Estas teorías pueden expresar una lógica de intervención de una política: las acciones políticas, al asignar ciertos recursos financieros tienen como objetivo producir unos productos y servicios planificados a través de los cuales se espera lograr los resultados esperados en términos de bienestar y progreso de las personas. Los resultados reales dependerán tanto de la efectividad de la política como de otros factores que afecten a los resultados, incluido el contexto. Sin duda, un elemento esencial de la efectividad de las políticas son los mecanismos que hacen que la intervención funcione.

La evaluación basada en la teoría explora los mecanismos que los responsables de políticas creen que hacen que la política sea efectiva y los compara con la evidencia basada en la investigación. Por lo tanto, su objetivo es encontrar y articular esta teoría, probarla y mejorarla, si es necesario.

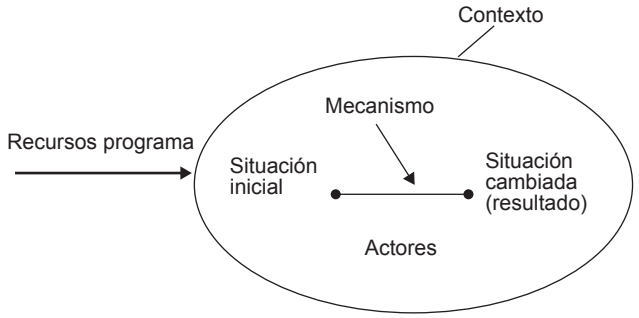
Este tipo de evaluación tiene dos componentes. El primero es conceptual, el segundo empírico. Conceptualmente, las evaluaciones articulan una teoría de política o programa, mientras que a nivel empírico, las evaluaciones buscan probar esta teoría para investigar si, por qué o cómo las políticas o los programas producen resultados observados en comparación a los planificados.

Evaluación realista

La evaluación realista es una aplicación del realismo científico a la evaluación. El realismo científico (Bhaskar, 2009) enmarca la realidad como un elemento estratificado hecho de capas anidadas, donde la intervención pública que se desea estudiar está totalmente integrada y, como tal, depende del contexto.

ESQUEMA 2

CONFIGURACIÓN EN UNA EVALUACIÓN REALISTA



FUENTE: Elaboración propia.

El mensaje básico de la evaluación realista es que la evaluación debe centrarse en comprender qué funciona mejor para quién, en qué circunstancias y, en particular, qué elementos de un programa hacen que la intervención tenga éxito. Para hacerlo, necesita desentrañar los «mecanismos internos» que funcionan en diferentes contextos, porque las intervenciones no tienen el mismo grado de eficacia en todas partes.

Técnicamente, la evaluación realista implica la identificación de una o más configuraciones de contexto-mecanismo-resultados (CMO), donde los contextos están hechos de recursos, oportunidades y limitaciones disponibles para los beneficiarios; los mecanismos son opciones, razonamientos o decisiones que los individuos toman en función de los recursos disponibles en su contexto; y los resultados son el producto del comportamiento y las elecciones de los individuos. Las configuraciones de CMO a menudo se representan con el Esquema 2.

Los diferentes contextos en los que opera el programa suponen una diferencia en los resultados que logra dicha intervención. Estos contextos incluyen características tales como estructuras sociales, económicas y políticas, contexto organizacional, partes interesadas del programa, personal del programa, contexto geográfico e histórico, etc. Algunos factores en el contexto pueden permitir la activación de mecanismos particulares del programa, mientras que otros aspectos pueden evitar que se activen estos mecanismos. El objetivo de la evaluación es aprender más sobre «qué funciona para quién», «en qué contextos los programas particulares no funcionan» y «qué mecanismos son activados por qué programas en qué contextos» (Westhorp, 2014).

Análisis comparativo cualitativo

El análisis comparativo cualitativo (ACC) es un método para la comparación sistemática de diferentes casos, y fue propuesto por Charles Ragin en 1987 (Ragin, 1987) para determinar qué factores cualitativos pueden influir en el resultado de un programa público. En esencia, el ACC requiere la conceptualización de casos (p. ej., proyectos o grupos de proyectos) como combinaciones o «paquetes» de características que se sospecha que influyen causalmente en un resultado. Por ejemplo, se supone que la disponibilidad de piezas de repuesto y mano de obra adecuadamente capacitada influye en la posibilidad de reparar fugas y filtraciones en sistemas de canalización de agua

(Welle, Williams, Pearce y Befani, 2015). Estas características del «caso» se denominan «condiciones» en lugar de «variables».

Una vez que se conocen las condiciones (características) de los casos, junto con sus resultados, se lleva a cabo una comparación sistemática entre casos para verificar qué factores se asocian sistemáticamente con un determinado tipo de resultado, por ejemplo el éxito de la intervención, y pueden ser considerados causalmente responsables de ello.

A diferencia de la mayoría de los métodos destinados a extraer lecciones generalizadas en todos los casos, el ACC no mira las variables de forma aislada. Se enfoca en combinaciones o configuraciones de factores dentro de casos individuales; y permite la generalización solo en la medida en que se conserven estas combinaciones.

Aunque el ACC establece una asociación entre una condición «dependiente» (el resultado) y una serie de condiciones «independientes», el objetivo de la ACC no es medir la correlación, o comprender cuánto se «agrega» una variable dada al resultado, sino más bien establecer: *i*) cuáles son las condiciones necesarias para un resultado y *ii*) cuáles son las combinaciones suficientes de condiciones para lograr el mismo resultado. La «necesidad» causal significa que se requiere un resultado: nunca se puede observar ese resultado sin la presencia de ciertas condiciones, mientras que «suficiencia» significa que la combinación es lo suficientemente buena para producir el resultado y no necesita ningún otro requisito.

Análisis de contribución

El análisis de contribución (AC) (Mayne, 2001 y 2008) se basa en un análisis en detalle de la teoría del cambio de la intervención que se evalúa. Dependiendo de la situación, esta teoría del cambio puede basarse en las expectativas de los financiadores, la comprensión de quienes gestionan la intervención, las experiencias de los beneficiarios y/o los resultados de investigaciones y evaluaciones anteriores. Esta teoría del cambio puede desarrollarse durante la planificación de la intervención, y luego revisarse a medida que se produce la implementación, o puede construirse retrospectivamente en el momento de una evaluación, y en ambos casos resulta una buena práctica el aprovechar investigaciones anteriores sobre intervenciones similares. El análisis realizado examina y prueba la teoría del cambio contra la lógica y los datos disponibles de los resultados observados, así como los diversos supuestos detrás de la teoría del cambio, y examina otros factores que influyen. El análisis confirma la teoría del cambio postulada o sugiere revisiones en las que la realidad parece ser diferente. El objetivo general es reducir la incertidumbre acerca de la contribución que una intervención está haciendo a los resultados observados, a través de una mayor comprensión de por qué los resultados ocurrieron o no y los roles desempeñados por la intervención y otros factores que influyen.

Los seis pasos clave para emprender un AC son: *i*) establecer el problema de causa-efecto que se abordará; *ii*) desarrollar la teoría del cambio y los riesgos que se postulan, incluyendo otros factores que influyen; *iii*) reunir la evidencia existente sobre la teoría del cambio; *iv*) recopilar y evaluar la contribución, y posibles desafíos a este diseño; *v*) recopilar nueva evidencia de la implementación de la intervención; y *vi*) revisar y fortalecer la historia de la contribución.

Evaluación con contrafactual

Un enfoque contrafactual implica contestar a la pregunta: «¿en qué medida se puede atribuir un impacto específico a la intervención?», es decir, determinar la «atribución» de la intervención pública en la variable de resultado analizada. Para ello, se realiza una estimación de lo que habría sucedido en ausencia de un programa o política —grupo de control— y se compara esto con lo que se ha observado en presencia de la intervención —grupo de tratamiento— (Pérez y Moral-Arce, 2015

y Khandker, Koolwal y Samad, 2009). En el siguiente apartado se desarrolla este tipo de evaluaciones[4].

7. Diseño de contrafactuales en evaluación de impacto

El impacto del programa hay que entenderlo como una diferencia en el valor de resultado específico (p. ej., salario, peso, puntuación en un examen, etc.) para el mismo individuo cuando recibe el programa y cuando no lo recibe (en condiciones *ceteris paribus*). Esta definición es válida para cualquier unidad analizada de observación, ya sean personas, granja, empresa, comunidad, pueblo, región, área de programación, país, etc., y cualquier variable de resultado que se pueda relacionar con el programa analizado. En esta situación, el reto de la evaluación es determinar los efectos del programa en condiciones reales, es decir, calcular el alcance y la magnitud de los cambios en la variable de resultado que se deben únicamente a la intervención política y no a otros factores. Para poder contestar a la pregunta de El planteada inicialmente y aislar el efecto exclusivo de la política estudiada es necesario disponer de una expresión matemática de la cuestión planteada inicialmente: ¿cuál es la situación (en la variable de resultado) que hubiera tenido un beneficiario de una intervención pública en el caso de no ser beneficiario? Analizar el efecto que una política *D* ha tenido sobre la variable de resultado *Y*, matemáticamente se expresa del siguiente modo:

$$\alpha_i=(Y_i \mid D_i=1)-(Y_i \mid D_i=0) \qquad [1]$$

El impacto de una política *D* en un individuo *i*, que denominamos α_i, se puede expresar como la diferencia entre dos situaciones: el resultado que ha tenido un individuo si recibe la política, $(Y_i \mid D_i=1)$, menos el resultado que hubiera tenido ese mismo individuo en el caso de no recibir la política, $(Y_i \mid D_i=0)$. Esta expresión permite establecer perfectamente el efecto o impacto de un programa sobre un individuo, porque en ambas situaciones los individuos tienen los mismos factores, salvo un cambio: en uno de ellos recibe el programa mientras que en el otro no, por lo que la diferencia entre ambas situaciones solo puede ser debida al programa estudiado. Se ha conseguido aislar el impacto de la intervención pública que se buscaba inicialmente.

Lamentablemente, no es posible disponer de información de una persona en dos estados diferentes. Si la persona recibe la ayuda no se puede ver qué hubiera pasado en el caso de no recibirla, porque la recibió, y viceversa. Por lo tanto, la ecuación [1] no se puede utilizar en la realidad porque uno de los dos elementos de la parte derecha de la igualdad no es conocido por el evaluador. Es lo que en la literatura de evaluación se conoce como el «problema del contrafactual».

El problema del contrafactual

Para solucionar esta situación dada en [1], si se dispone de información de un individuo que ha recibido la ayuda, se propone buscar otro individuo que presente características iguales salvo que esta unidad no ha recibido la ayuda (Gertler, Martínez, Premand, Rawlings y Vermeersch, 2011).

Considerando la dificultad de tener dos individuos que sean exactamente iguales en todas sus características, salvo por el hecho de que uno recibe el programa (D=1) mientras que el otro no (D=0), la solución consiste en utilizar «grupos de individuos» en lugar de un solo individuo. Es muy probable que comparando persona a persona siempre exista un factor diferencial (además de recibir el programa), lo que

[4] Existen diferentes organismos que realizan evaluaciones con contrafactual en España, aunque se encuentran de manera dispersa, dado que no existe ningún centro que centralice y promueva este tipo de evaluaciones. En el ámbito de la Administración Pública, hay organismos como el Instituto de Estudios Fiscales (IEF) que colabora con FEDER entre otros ministerios, a departamentos centrados en las políticas y programas que desarrollan sus instituciones como son ICEX, CDTI o el Ministerio de Empleo en el Fondo Social Europeo. Fuera del ámbito del sector público, la academia produce este tipo de estudios de evaluación de impacto y también organismos como FEDEA.

hace imposible aislar el efecto del programa analizado. Sin embargo, al utilizar grupos de individuos, siempre se puede seleccionar una serie de personas que, en promedio, presenten características similares. Consideremos la siguiente situación: después de implementar este programa *D* en la población, existen dos tipos de individuos, los que han recibido el programa (D=1) y los que no (D=0). Con la información de los individuos de cada grupo, el impacto de una política se obtiene como:

$$\alpha=E(Y|\,D=1)-E(Y|\,D=0) \qquad [2]$$

Donde el término *E(.)* se refiere a la esperanza (o la media). Ahora, el impacto del programa se obtiene calculando el promedio de la variable de resultado en el grupo de individuos tratados, *E(Y|D=1)*, menos el resultado promedio obtenido a partir de los individuos que no han recibido el programa, *E(Y | D=0)*.

De forma similar al caso anterior, para aislar el efecto del programa, es decir, aquella parte del cambio en *Y* que se debe solo al programa y no a otros factores, es necesario que esas otras características sean similares, en media, en los dos grupos.

Sin embargo, la ecuación [2] no contesta exactamente a la pregunta de EI sobre la situación que hubiera tenido un individuo tratado en el caso de no haberlo sido. En términos matemáticos, esta pregunta viene dada por la siguiente expresión:

$$\alpha=E(Y_1-Y_0|\,D=1) \qquad [3]$$

La lectura de la ecuación es: para un beneficiario, *D=1*, ¿cuál es la diferencia en su variable de resultado, *Y*, entre haber recibido el tratamiento, $E(Y_1|\,D=1)$, respecto a la situación de no haberlo recibido -$E(Y_0|\,D=1)$? que es exactamente idéntico a la pregunta de EI. Deshaciendo el paréntesis la ecuación anterior se puede expresar como:

$$\alpha=E(Y_1|\,D=1)-E(Y_0|\,D=1) \qquad [4]$$

De manera similar a la ecuación [1], en esta expresión anterior existe un componente que nunca se puede obtener: $E(Y_0|\,D=1)$, porque es desconocido. Dado que estamos con personas que recibieron tratamiento, *D*=1, la pregunta ¿cuál hubiera sido el valor del resultado de un individuo tratado, en el caso de no haberlo sido? no tiene respuesta, porque todas recibieron tratamiento.

En evaluación de impacto, como el evaluador no puede obtener esta cantidad indicada previamente, lo que hace es buscar un grupo de individuos que no han recibido el programa, es decir, que en la realidad no son beneficiarios –*D*=0, y se ve la variable de resultados, *Y*. Dado que no son beneficiarios, la cantidad que se obtiene es $E(Y_0|\,D=0)$. Esto permite pasar de la ecuación [4] a la dada en [2], que sí se puede estimar.

Para que se pueda producir el paso de [4] a [2], el gran supuesto de EI es asumir que $E(Y_0|\,D=1) = E(Y_0|\,D=0)$ y, por lo tanto, de la ecuación [4] se sustituye la última parte por $E(Y_0|\,D=0)$. Entonces, el impacto que estima el evaluador es:

$$dif_{tra\text{-}cont}= E\,(Y_1|\,D=1) - E(Y_0|\,D=0) \qquad [5]$$

Lamentablemente, la ecuación que contesta al «impacto del programa» dada por [4] y lo que el evaluador puede calcular, dado por [5], no es exactamente lo mismo. Si a la expresión [4] se suma y resta la cantidad $E(Y_0|\,D=1)$ se obtiene la ecuación:

$$dif_{tra\text{-}cont}=E(Y_1|\,D=1)-E(Y_0|\,D=0)+ E\,(Y_0|\,D=1)-E(Y_0|\,D=1) \qquad [6]$$

Y reordenando términos se obtiene:

$$dif_{tra\text{-}cont}= \underbrace{\{E(Y_1|\,D=1) - E(Y_0|\,D=1)\}}_{\alpha} + \underbrace{\{E(Y_0|\,D=1)- E(Y_0|\,D=0)\}}_{\text{sesgo de selección}} \qquad [7]$$

Donde $\alpha=\{E(Y_1|\,D=1)–E(Y_0|\,D=1)\}$ *es* el impacto verdadero del programa, mientras que $\{E(Y_0|\,D=1)–E(Y_0|\,D=0)\}$ es lo que se denomina sesgo de selección. Así que la ecuación [7] se puede expresar como:

$$dif_{tra\text{-}cont}=\alpha+sesgo.selec \quad [8]$$

Por lo tanto, la estimación del impacto de un programa que calcula el evaluador es igual al verdadero impacto del programa, solo si el sesgo de selección es cero.

El sesgo de selección

Como se ha indicado, el supuesto necesario para que el impacto estimado por el evaluador sea igual al verdadero impacto del programa, es decir, que [4] sea igual a [5], es que se cumpla la siguiente hipótesis:

$$E(Y_0 \mid D=1)=E(Y_0 \mid D=0) \quad [9]$$

Esta expresión indica que un beneficiario, D=1, en el caso de no haber recibido la ayuda, $E(Y_0 \mid D=1)$, se debería haber comportado igual que una persona que en realidad nunca recibió ayuda, $E(Y_0 \mid D=0)$. En el caso de no cumplirse la ecuación [9] implica la aparición de sesgos de selección (Heckman, Lalonde y Smith, 1999).

Este sesgo de selección es la diferencia existente entre el resultado medio de no recibir tratamiento entre el grupo de tratamiento y el grupo de control, que además de ser el efecto del programa evaluado, puede ser por otras diferencias entre las personas tratadas y el grupo de control que no solo afecten a la participación en el programa (ser D=1 o D=0), sino también a la variable de resultado.

Una forma de entender el sesgo de selección consiste en considerar que los individuos presentan determinadas características —edad, sexo, nivel de estudios, gustos— que hacen que sean más propensos a pertenecer a un grupo de tratamiento o al grupo de control. Por lo tanto, que un individuo esté en el grupo de tratamiento no es aleatorio, y no solo eso, sino que esa característica que hace que sea más proclive a pertenecer al grupo de tratamiento también puede afectar a la variable de resultado. Los diferentes tipos de sesgos de selección son: *i*) sesgo establecido por la política (focalización); *ii*) sesgo de los potenciales beneficiarios de la política, que se autoseleccionan; *iii*) sesgo en variable observada: si la característica que afecta a que un individuo esté en el grupo de tratamiento o control se encuentra en el fichero de información que posee el evaluador; y *iv*) sesgo en variable no observada, si esa característica no está disponible para el evaluador.

Teniendo en cuenta la ecuación que combina el impacto del programa y la estimación que realiza el evaluador, el elemento fundamental para lograr buenas estimaciones está en cancelar el sesgo de selección dado en [8]. Para lograr que la cantidad que calcula el evaluador, $dif_{tra\text{-}cont}$, sea igual al impacto verdadero del programa, α, existen dos posibilidades: *i*) «diseños experimentales» en los que la asignación de los individuos al grupo de control y tratamiento se realiza de forma aleatoria; y *ii*) «diseños cuasi experimentales».

Diseños experimentales

Las características necesarias para que exista un «diseño aleatorio o experimental» son las siguientes: *i*) la evaluación se diseña a la vez que la política; *ii*) el evaluador determina al «azar» qué individuo está en el grupo de control y tratamiento, por ejemplo, mediante el lanzamiento de un dado; y *iii*) una vez que un individuo ha sido seleccionado en un grupo, nadie puede cambiar del grupo donde ha caído (Duflo y Banerjee, 2017).

A continuación, se implementa la política solo sobre aquellos individuos que fueron seleccionados —aleatoriamente— en el grupo de tratamiento. Pasado un tiempo, se observa la variable de resultado. Esta aproximación hace que el sesgo de selección sea cero.

Diseño experimental: diferencia de medias

Tras determinar de forma aleatoria quién sí (no) recibe el programa, a continuación se ejecuta la

intervención pública a aquellos individuos que fueron seleccionados en el grupo de tratamiento. En el período posterior se obtiene información de la variable de resultado de los individuos de los dos grupos. La El implica calcular la diferencia entre dos grupos, dada por la siguiente ecuación:

$$\alpha=E(Y\mid D=1)-E(Y\mid D=0)=\bar{Y}_D-\bar{Y}_C \qquad [10]$$

Sin embargo, la cuantía calculada no permite determinar si es lo suficientemente grande para decir que hay impacto. Por ese motivo, para contestar a «si existe impacto» es necesario realizar un contraste de hipótesis, donde:

Hipótesis nula: el programa no tiene impacto (no es efectivo) → H_0:$\bar{Y}_D=\bar{Y}_C$

Hipótesis alternativa: el programa sí tiene impacto (es efectivo) → H_A:$\bar{Y}_D\neq\bar{Y}_C$

La lectura de la hipótesis nula es la siguiente: la media de la variable de resultado es estadísticamente igual a la media de la variable del grupo de control. Esto es similar a escribir que «no se observan diferencias en la variable de resultado entre los dos grupos», y que por lo tanto se asume que el programa no tiene ningún impacto.

Para contrastar este conjunto de hipótesis, el estadístico que se calcula es:

$$t_0=\frac{\bar{Y}_D-\bar{Y}_C}{\sqrt{\frac{S_D^2}{N_D}+\frac{S_C^2}{N_C}}} \qquad [11]$$

donde S_D^2, S_C^2 son la varianza del grupo de tratamiento y control respectivamente, mientras que N_D, N_C son el total de individuos que hay en el grupo de tratamiento y grupo de control. En el caso de $|t_0|\geq 1{,}96$ entonces se rechaza la hipótesis nula (a un nivel de significatividad del 5 %), y se considera que el programa evaluado tiene impacto, en términos estadísticos, y la cuantía de dicho impacto será la dada por α.

Diseños cuasi experimentales

Cuando alguna de las tres condiciones necesarias para tener un diseño experimental no se cumple entonces nos encontramos ante un diseño cuasi experimental. Ahora, lo que estima el evaluador no coincide con el impacto verdadero, así que será necesario utilizar técnicas estadísticas y econométricas para lograr minimizar lo máximo posible esa cuantía del sesgo y hacerla cero. Métodos habituales en este entorno son la técnica de emparejamiento, diferencias en diferencias o regresión en discontinuidad, entre otros.

Diseño cuasi experimental: diferencia en diferencias

Cuando el sesgo de selección de los individuos a participar en el grupo de tratamiento depende de variables no observadas, el método de diferencias en diferencias, también conocido como *dif-in-dif*, ofrece una solución de corrección del sesgo, permitiendo obtener estimaciones del impacto de una intervención pública con buenas propiedades.

Para poder utilizar el método de *dif-in-dif* es necesario disponer de información de la variable de resultado *Y* en dos momentos de tiempo: la línea de base (momento *t=0*) y posterior a tratamiento (momento *t=1*).

Si se dispone de información del grupo de control y tratamiento en estos dos períodos, antes y después de que se produzca la intervención pública, el método consistirá en calcular la evolución temporal de la variable *Y* de cada grupo (primera diferencia), y con posterioridad ver el diferencial en el crecimiento que se produjo entre el control y el tratamiento (segunda diferencia). Las fases de estimación son las siguientes:

1ª etapa: estimación del promedio de la variable de interés en cuatro situaciones:

$E(Y \mid tratado, antes) = \bar{Y}_{D0}$ $E(Y \mid tratado, después) = \bar{Y}_{D1}$

$E(Y \mid control, antes) = \bar{Y}_{C0}$ $E(Y \mid control, después) = \bar{Y}_{C1}$

2ª etapa: 1ª diferencia: estimación de la diferencia temporal para cada grupo:

Grupo de tratamiento: $\Delta Y_D = Y_{D1} - Y_{D0}$ [12]

Grupo de control: $\Delta Y_C = Y_{C1} - Y_{C0}$ [13]

Estas diferencias indican cuánto ha crecido, en promedio, la variable de interés durante los dos períodos de tiempo, de *t0* a *t1*, para cada uno de los grupos analizados.

3ª etapa: diferencia de la diferencia: cálculo del diferencial entre los dos grupos de la diferencia temporal obtenida en la etapa anterior:

$$\alpha = \Delta Y_T - \Delta Y_C = (\bar{Y}_{D1} - \bar{Y}_{D0}) - (\bar{Y}_{C1} - \bar{Y}_{C0}) \quad [14]$$

La ecuación anterior nos ofrece la «cuantía» del impacto, pero por sí sola no es capaz de informar si el programa tiene impacto estadísticamente significativo sobre la variable de interés, así que resulta necesario realizar un contraste de hipótesis:

Hipótesis nula: el programa no tiene impacto $\rightarrow H_0: \Delta \bar{Y}_D = \Delta \bar{Y}_C$

Hipótesis alternativa: el programa sí tiene impacto $\rightarrow H_A: \Delta \bar{Y}_D \neq \Delta \bar{Y}_C$

Donde $\Delta \bar{Y}_D = (\bar{Y}_{D1} - \bar{Y}_{D0})$ y $\Delta \bar{Y}_C = (\bar{Y}_{C1} - \bar{Y}_{C0})$, respectivamente. El estadístico que se calcula en el contraste de hipótesis es:

$$t_0 = \frac{\Delta \bar{Y}_D - \Delta \bar{Y}_C}{\sqrt{\frac{S_D^2}{N_D} + \frac{S_C^2}{N_C}}} \quad [15]$$

donde S_D^2, S_C^2 son la varianza del grupo de la variable $\Delta \bar{Y}$ del grupo de tratamiento y control respectivamente y *N* el tamaño de muestra de cada grupo. En el caso de $|t_0| \geq 1{,}96$ entonces se rechaza la hipótesis nula, y se considera que el programa evaluado tiene impacto, en términos estadísticos, y que la cuantía de dicho impacto será la dada por *α* (Card y Krueger, 1994).

Diseño cuasi experimental: método de emparejamiento

En la mayoría de intervenciones públicas de las que se realiza una evaluación de impacto es posible que existan sesgos de selección en variables observadas.

Este método consiste en emparejar individuos, tratados y no tratados, en función de las características de estos individuos. Sin embargo, a la hora de emparejar no se tiene en cuenta variables observadas como edad, comunidad autónoma, nivel de estudios, rama de actividad, etc., sino que mediante la aplicación de técnicas cuantitativas todas estas posibles combinaciones de características se colapsan en una única característica que indique la probabilidad que un individuo tiene de pertenecer al grupo de tratamiento (*D*=1) de acuerdo a sus características observadas. Por lo tanto, ahora el emparejamiento de individuos para calcular el impacto será en función de esta propensión, o probabilidad que tengan para ser beneficiarios (Rosenbaum y Rubin, 1983; Caliendo y Kopeinig, 2008). A continuación se describen las diferentes etapas para aplicar este método:

a) Estimar el *propensity score*. No es más que una regresión econométrica donde la variable dependiente es «participar o no en el programa, D_i» en función de las características de los individuos, X_i. Estos modelos, como la variable dependiente, solo pueden tomar dos valores (0,1) y se estiman con un modelo probit o logit. La especificación econométrica de este último viene dada por:

$$Pr(D_i = 1 \mid X_i) = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \quad [16]$$

Una vez estimados los parámetros $\hat{\beta}_o$, $\hat{\beta}_1$, se puede calcular, para cada unidad de la muestra de datos analizada, la probabilidad de un individuo de participar en el programa, de acuerdo a sus características observadas X_i, es decir, $\widehat{Pr}(D_i=1|X_i)$.

b) Se contrasta si existen diferencias entre los grupos de tratamiento y de control para diferentes niveles de «propensión a participar», mediante la realización de contrastes de igualdad de medias entre los grupos de tratamiento y control en las variables explicativas *X*, también conocido como «test de equilibrado» o *balancing test*.

c) Una vez determinada la especificación, las observaciones se ordenan en forma ascendente de acuerdo con el valor estimado de la propensión a participar en el programa (*propensity score*), descartándose aquellas observaciones de «no tratados» con valores estimados del *propensity score* estimado demasiado extremos.

d) Se procede a emparejar a cada individuo tratado con aquel individuo, o grupo de individuos, no tratado cuyo valor estimado del *propensity score* esté más próximo.

e) Se compara la variable de resultado *Y* entre los individuos de tratamiento y control que se consideran similares y, finalmente, se obtiene la media de estas diferencias individuales para obtener el impacto medio del programa:

$$\hat{\alpha} = \frac{1}{N_D}\sum_{i\in D}[Y_i^D - \sum_{j\in C} w(i,j)Y_j^C] \qquad [17]$$

Donde Y_i^D es el valor de la variable de resultado del tratado, mientras que Y_j^C es la variable de resultado del tratado «*j*», siendo *w(i,j)* una funcion de peso que determina cómo ponderar los controles empleados en este cálculo, y N_D el total de individuos tratados que se han empleado en el cálculo.

8. Conclusiones

En este trabajo se realiza una presentación de la evaluación *ex post* de una intervención pública, comenzando desde los análisis más cercanos a gestión que estudian el uso de recursos y los productos logrados, pasando a exponer aproximaciones centradas en el análisis de los resultados obtenidos por esa intervención pública. Dentro de este enfoque, destaca la evaluación de impacto, que pretende valorar si un programa logra el resultado propuesto inicialmente. A continuación se muestran las diferentes preguntas de evaluación de impacto que se pueden realizar y las metodologías empleadas para contestar a cada una de estas preguntas. Se finaliza el trabajo desarrollando las evaluaciones con contrafactuales, donde el sesgo de selección es un elemento de gran importancia, y los diferentes tipos de diseños a emplear permitan una mejor estimación del impacto del programa analizado.

Referencias bibliográficas

Befani, B. (2016). *Choosing Appropriate Evaluation Methods. A tool for assessment and selection*. London, UK: Bond.

Bhaskar, R. (2009). *Scientific realism and human emancipation*. Routledge.

Caliendo, M. y Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22, 31-72.

Card, D. y Krueger, A. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review,* (84), 772-793.

Duflo, E. y Banerjee, A. (2017). *Handbook of Field Experiments,* 1. North Holland.

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B. y Vermeersch, C. M. (2011*). L'évaluation d'impact en pratique*. The World Bank.

Heckman, J. J., Lalonde, R. y Smith, J. (1999). The economics and econometrics of active labor market programs. En Ashelfelter, O. y Card, D. (Eds). *In Handbook of Labor Economics*, 3, Amsterdam: Elsevier.

Imas, L. G. y Rist, R. (2009*). The road to results: Designing and conducting effective development evaluations*. Washington, DC: The World Bank.

Khandker, S. B., Koolwal, G. B. y Samad, H. A. (2009*). Handbook on impact evaluation: quantitative methods and practices*. The World Bank.

Kirkpatrick, D. (1995). *Evaluating training programmes*. San Francisco: Berrett-Koehler.

Mayne, J. (2001). Addressing attribution through contribution analysis: using performance measures sensibly. *The Canadian Journal of Program Evaluation*, 16, 1-24.

Mayne, J. (2008). Contribution analysis: an approach to exploring cause and effect. ILAC Brief (16). Institutional learning and change (ILAC) initiative (CGIAR).

Moral-Arce, I., Paniagua, M., Rodríguez, L. y Rodríguez, C. (2016). *Evaluación de políticas públicas: Técnicas cuantitativas*. Madrid: Garceta grupo editorial.

Pérez, C. y Moral-Arce, I. (2015). *Técnicas de evaluación de impacto.* Madrid: Garceta grupo editorial.

Ragin, C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley, Los Angeles & London: University of California Press.

Rosenbaum, P. y Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-50.

Stern, E. (2015). Impact Evaluation: *A guide for commissioners and managers*. Prepared for the Big Lottery Fund, UK: Bond, Comic Relief and the Department for International Development.

Welle, K., Williams, J., Pearce, J. y Befani, B. (2015). *Testing the waters: a qualitative comparative analysis of the factors affecting success in rendering water services sustainable based on ICT reporting.* Brighton: Institute of Development Studies and WaterAid.

Westhorp, G. (2014). Realist impact evaluation: an introduction. London: Overseas Development Institute.