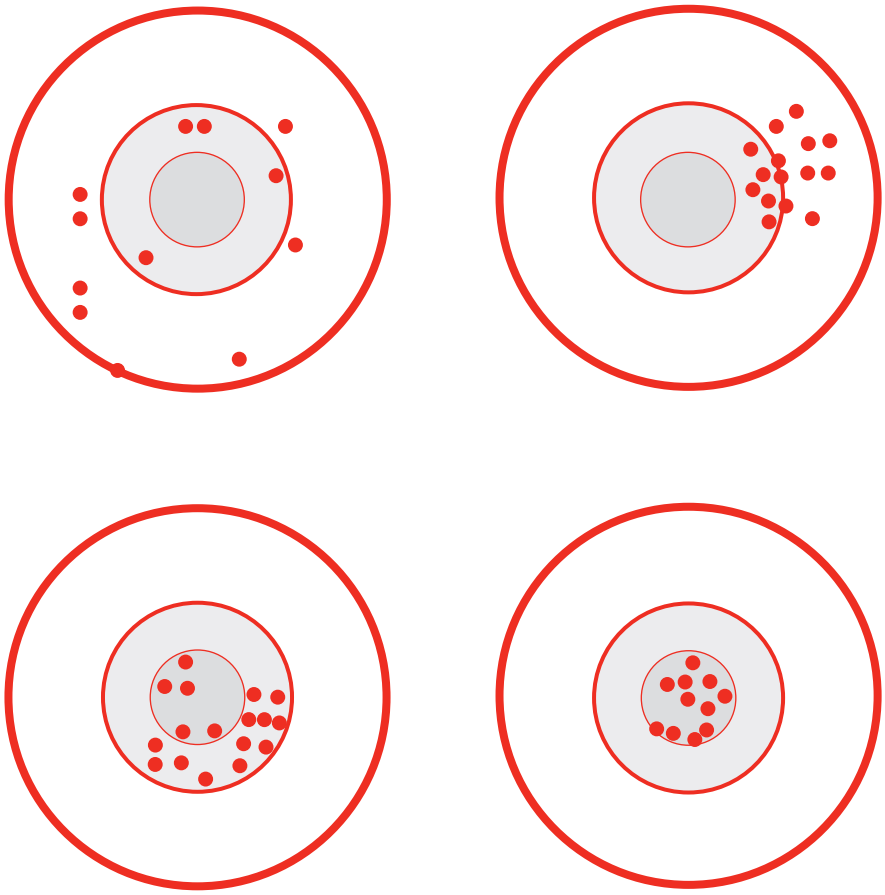
Guía práctica 5
Evaluación de impacto

Colección Ivàlua de guías prácticas sobre evaluación de políticas públicas



ivàlua | Institut Català d'Avaluació de Polítiques Públiques

Instituciones miembros de Ivàlua:

Generalitat de Catalunya Departament d'Economia i Finances

Diputació Barcelona

Universitat Pompeu Fabra

Fundació Jaume Bofill

Autores: Jaume Blasco, analista de Ivàlua
David Casado, analista de Ivàlua
Diseño: petitcomite.net
Impresión: Cevagraf, s.c.c.l.

Primera edición: diciembre de 2009

Dipósito legal: B-45840-2009

ÍNDICE

1. INTRODUCCIÓN PÁG. 5

1.1. EVALUACIÓN DE IMPACTO: EN BUSCA DE LA CAUSALIDAD pág. 6
1.2. EL CONTRAFACTUAL Y LA ESTIMACIÓN DEL IMPACTO DE UNA POLÍTICA PÚBLICA pág. 9

2. PASOS PRELIMINARES PARA DISEÑAR UNA EVALUACIÓN DE IMPACTO PÁG. 13

2.1. ¿ES OPORTUNO EVALUAR LOS IMPACTOS DEL PROGRAMA? pág. 13
2.2. ¿A QUÉ NOS REFERIMOS CUANDO HABLAMOS DE *OUTCOMES*? pág. 14
2.3. ¿QUÉ QUIERE DECIR PARTICIPAR EN EL PROGRAMA? pág. 18
2.4. ¿PARA QUIÉN QUEREMOS DETECTAR LOS IMPACTOS? pág. 19
2.5. ¿A QUÉ NOS REFERIMOS EXACTAMENTE CUANDO HABLAMOS DE CONTRAFACTUAL? pág. 20
2.6. ¿DE QUÉ DATOS DISPONEMOS PARA HACER LA EVALUACIÓN DE IMPACTO? pág. 21

3. MÉTODOS PARA LA EVALUACIÓN DE IMPACTO PÁG. 23

3.1. LA VALIDEZ DE LAS CONCLUSIONES pág. 24
3.2. EXPERIMENTOS SOCIALES pág. 27
3.3. DISEÑO SIN GRUPO DE CONTROL: ANTES-DESPUÉS Y SERIES TEMPORALES pág. 34
3.4. LA TÉCNICA DEL *MATCHING* pág. 37
3.5. EL MODELO DE DOBLES DIFERENCIAS pág. 41
3.6. ELECCIÓN ENTRE MÉTODOS pág. 45

BIBLIOGRAFÍA PÁG. 51

ANEXO. GUÍA DE RECURSOS PÁG. 52

MANUALES pág. 52
ARTÍCULOS pág. 52
ENLACES DE INTERÉS pág. 54

1. INTRODUCCIÓN

Las administraciones públicas se dedican continuamente a diseñar e intentar mejorar políticas y programas, y dedican cada año miles de millones de euros a implementarlos. A pesar de ello, problemas como el desempleo, el fracaso escolar, la siniestralidad en las carreteras o la degradación ambiental tienden a persistir, lo cual plantea dudas sobre la efectividad de las intervenciones públicas que deben enfrentarse a ellos. Por una parte, este hecho pone de manifiesto que la tarea de enfrentarse a los problemas sociales es complicada, que en el mejor de los casos da lugar a avances lentos, graduales e incompletos. Por otra, que, aunque una intervención pública parezca una gran idea y se destinen a ella muchos recursos, su éxito no puede darse nunca por garantizado *a priori*.

Sobre la base de un análisis sistemático *ex post*, **la evaluación de impacto** trata, precisamente, de determinar la capacidad que tienen las ideas potencialmente buenas para solucionar los problemas sociales. ¿Un aumento de los impuestos sobre el tabaco consigue realmente que la gente fume menos? ¿Ofrecer desgravaciones fiscales para los planes de pensiones consigue que la gente ahorre más para después de la jubilación? ¿Incrementar las horas lectivas en la educación primaria mejora el rendimiento escolar? ¿Formar a desempleados con baja cualificación aumenta su renta a medio plazo? Dado que los problemas sociales pueden acarrear consecuencias graves para quien los sufre y que los recursos para abordarlos son limitados, se trata de identificar y distinguir las políticas públicas que mejor consiguen solucionarlos o, como mínimo, contenerlos.

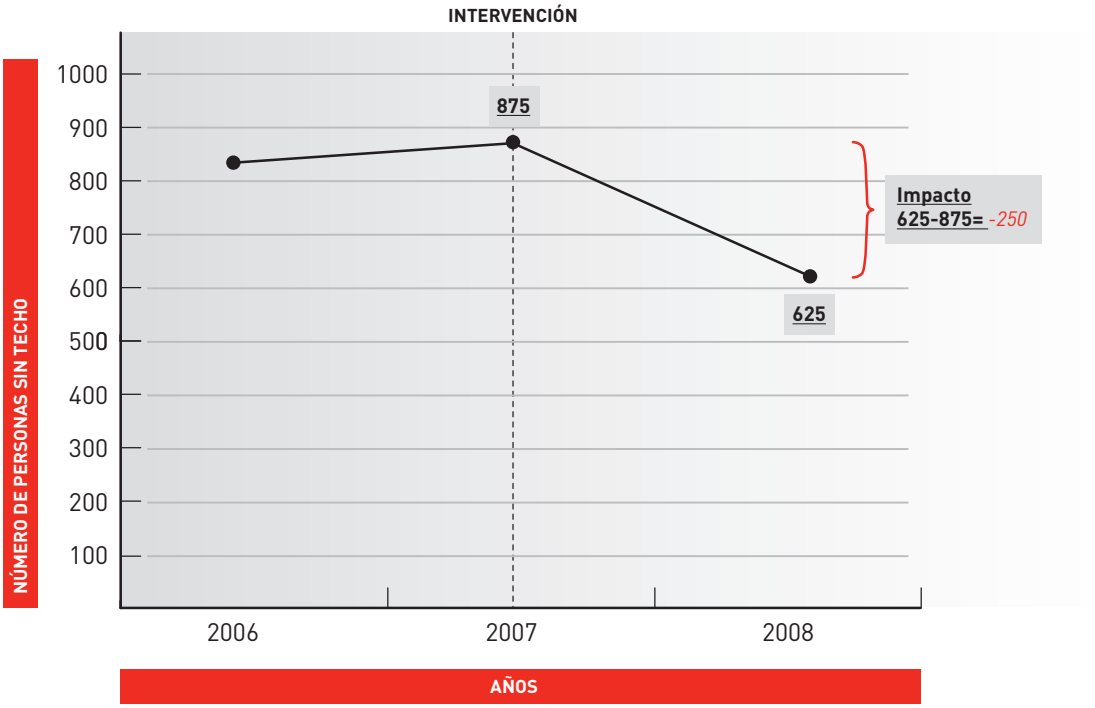
Pero ¿cómo podemos saber si las políticas públicas realmente funcionan? Y si funcionan, ¿cuál es la magnitud de su impacto? Demasiado a menudo la evaluación de las políticas se ha centrado exclusivamente en los *inputs* o los *outputs*, es decir, en los recursos que utiliza el programa o en aquello que hace el programa. Sin embargo, el que 100 bomberos hayan estado 24 horas echando agua sobre un fuego nos dice poco sobre si han conseguido apagarlo. Tampoco el mero seguimiento de un problema nos dice demasiado sobre el impacto real de las políticas públicas. Que un año se hayan quemado la mitad de hectáreas de bosque que el año anterior no quiere decir, necesariamente, que los bomberos hayan hecho mejor su trabajo. Por tanto, en una evaluación de impacto no solamente queremos saber si un problema mejora o empeora, sino si la intervención pública ha tenido algo que ver. Se trata, en resumen, de establecer si puede atribuirse o no (y en qué medida) la *causa* del cambio en el problema a la intervención pública. Hacerlo de forma convincente, como veremos, es una tarea laboriosa.

1.1. EVALUACIÓN DE IMPACTO: EN BUSCA DE LA CAUSALIDAD

Supongamos que el ayuntamiento de una ciudad pone en marcha un programa nuevo de atención a las personas sin techo que pernoctan en las calles. La intervención consiste en ofrecer atención personalizada en la calle a las personas que no utilizan los recursos municipales residenciales, con el propósito de que cada persona reciba siempre la atención del mismo trabajador social. El programa es costoso, ya que implica la contratación de numerosos trabajadores sociales nuevos, pero se espera que ayude a reducir considerablemente el número de personas que tienen que dormir a la intemperie. La teoría consiste en que, mediante este tipo de atención, el trabajador social desarrollará una relación de confianza con la persona sin techo que le permitirá detectar mejor qué problemas sufre, informarla, orientarla y acompañarla al recurso o servicio más adecuado en cada caso, e ir venciendo las barreras y desconfianzas que provocan que las personas sin techo se muestren reticentes a utilizar estos recursos. Se supone que de esta manera se incrementará el número de personas que entran en contacto con el sistema de atención, lo cual constituye un primer paso crítico para poder proporcionarles la asistencia que necesitan y, en último término, permitir que estas personas accedan a soluciones residenciales estables en las que puedan desarrollar su proyecto de vida con la mayor autonomía personal posible. La idea parece buena, pero ¿funcionará en la práctica?

Imaginemos que este ayuntamiento, para medir el impacto de sus programas para las personas sin techo, realiza recuentos anuales de las personas que pernoctan en la calle. Tal y como muestra el gráfico 1, los recuentos realizados con anterioridad al nuevo programa indicaban que en el año 2007 había 875 personas durmiendo en las calles de la ciudad, lo que representaba una pequeña variación respecto al año anterior. El recuento del año 2008 —esperado con expectación para poder estimar el impacto del nuevo programa— revela que la población de personas sin techo ha bajado hasta las 625 personas. En otras palabras, esto implica una reducción de 250 personas, casi un 30 % comparado con la población sin techo del año anterior. A primera vista parece que el impacto del programa ha sido positivo y considerable. Sin embargo, ¿podemos considerar que esta conclusión es suficientemente exacta?

Gráfico 1. Evolución de las pernoctaciones en la calle



Para responder a esta pregunta debemos tener en cuenta que, entre los años 2007 y 2008, pueden haber sucedido otras cosas aparte de la puesta en marcha del nuevo programa. Por ejemplo, es posible que la economía haya crecido y ofrezca más oportunidades laborales incluso para las personas de más baja cualificación. También puede haber ocurrido que los servicios de salud mental hayan iniciado un nuevo programa en coordinación con los servicios sociales que se haya mostrado especialmente efectivo para prevenir que las personas con enfermedades mentales graves y pocos recursos económicos acaben en la calle. Igualmente, es posible que el Gobierno haya endurecido el control de entrada al país de nuevos inmigrantes, dificultando así la llegada de inmigrantes indocumentados, los cuales constituyen un sector de la población con problemas muy graves de acceso a la vivienda. Todos estos fenómenos, entre muchos otros, podrían explicar, total o parcialmente, el descenso de la población sin techo observado entre 2007 y 2008. La situación contraria es igualmente factible: que en este mismo año las condiciones económicas hubieran empeorado, que se hubiera suprimido un programa de atención a las personas con enfermedades mentales y que hubieran entrado en la ciudad muchos más inmigrantes indocumentados que en años anteriores. En este caso, la reducción de 250 personas respecto al año anterior estimada en el gráfico 1 sería una clara subestimación del impacto real del programa.

La situación descrita en el ejemplo es la más habitual en una evaluación de impacto. Podemos medir fácilmente un determinado fenómeno, como pueda ser la cantidad de personas que duermen en la calle, el número de accidentes en las carreteras o la productividad del sector de la fruta dulce, para capturar el impacto o *outcome* de una intervención pública que nos interesa evaluar. Pero, desafortunadamente para los evaluadores, suceden muchas otras

cosas más allá de la propia intervención pública (como la evolución de la economía, los cambios en la meteorología o la puesta en marcha de otros programas y políticas) que tienen una influencia notable sobre el impacto que intentamos observar y que complican la evaluación. Por consiguiente, evaluar el impacto de un programa implica ser capaz de aislar el efecto del programa en relación con todos estos otros fenómenos que afectan al problema o situación que la intervención pública pretende abordar.

Esta constatación nos lleva introducir lo que parece un pequeño matiz, pero que tiene en realidad una importancia crucial en la evaluación de impacto (y que, como veremos más adelante, es la principal fuente de quebraderos de cabeza metodológicos): la pregunta que la evaluación de impacto debe responder no es qué ha pasado después de poner en marcha una intervención pública (muchas cosas pueden haber influido), sino qué ha pasado *en comparación con lo que habría ocurrido si la intervención no se hubiera llevado a cabo*. Lógicamente, la diferencia entre lo que ha sucedido con el programa y lo que habría sucedido sin el programa puede atribuirse sola y únicamente al programa o, dicho de otro modo, la diferencia ha sido *causada* por el programa. Y esto es, precisamente, lo que busca la evaluación de impacto: lo que el programa ha causado, y no lo que ha sucedido al mismo tiempo que el programa.

CUADRO 1
ASOCIACIÓN NO QUIERE DECIR CAUSALIDAD

Una de las reglas de oro presentes en casi todos los manuales de estadística es no confundir asociación con causalidad. La diferencia entre ambos conceptos es sencilla. Supongamos que, en un momento dado, observásemos en una población determinada que el hecho de tener los dedos amarillentos y el hecho de sufrir bronquitis crónica están asociados, es decir, son características que tienden a presentarse juntas en las mismas personas. ¿Quiere esto decir que la bronquitis crónica hace que la gente tenga los dedos amarillos? En realidad sabemos que no es así, sino que existe un tercer factor, que es fumar, que es una causa importante tanto de que la gente tenga los dedos amarillos, como de que padezcan bronquitis crónica. Por eso tener bronquitis y los dedos amarillos son fenómenos asociados, pero uno no es la causa del otro. Técnicamente se dice que la asociación que existe entre ambos fenómenos es *espuria*.

Pero desenredar causalidad y asociación en el campo de las políticas públicas no siempre es tan sencillo. Imaginemos que, entre la población escolar, estudiar en un colegio concertado está asociado con un mejor rendimiento académico que hacerlo en un colegio público. ¿Quiere esto decir que la titularidad del colegio es la causa de esta diferencia y que, por tanto, el concierto escolar es una forma de provisión de la educación más efectiva que la gestión pública directa? Es posible, pero no es seguro. Una explicación alternativa es que los alumnos del colegio concertado tienden a pertenecer a familias de un nivel socioeconómico y formativo superior al de las del público, y que esta diferencia en las características del alumnado es la causa real de la diferencia en el rendimiento escolar. De forma similar, que un ayuntamiento ponga en marcha un programa de atención a las personas sin techo y al año siguiente baje considerablemente el número de personas que pernoctan en la calle son hechos asociados, pero no necesariamente uno es la causa del otro. Como hemos visto en la explicación del ejemplo, existen muchos otros motivos plausibles, de modo que es mejor no extraer conclusiones precipitadas que nos puedan convertir en víctimas de la *falacia causal*.

Cuando observamos una asociación (por ejemplo, que participar en un programa está asociado a una mejora en un determinado *outcome*), es importante tener siempre presente que la causalidad es una explicación posible, pero no la única. El reto de la evaluación de impacto es, justamente, descartar explicaciones alternativas para poder atribuir, de la forma más convincente posible, la causalidad del cambio observado a la intervención pública.

1.2. EL CONTRAFACTUAL Y LA ESTIMACIÓN DEL IMPACTO DE UNA POLÍTICA PÚBLICA

Siguiendo la argumentación del párrafo anterior, el impacto de la intervención pública puede expresarse en términos de la diferencia entre dos números:

$$\text{IMPACTO} = Y_1 - Y_0$$

Donde:

- Y_1 son los *outcomes* que han ocurrido con la intervención pública.
- Y_0 son los *outcomes* que se habrían dado en ausencia de la intervención pública, que de forma más técnica (y más breve) se denominan *contrafactual*.

Por regla general, Y_1 es un número relativamente fácil de estimar. Normalmente, utilizando registros administrativos, mediante una encuesta, realizando un recuento (como en el ejemplo) o con cualquier otra técnica de observación, podemos estimar qué ha pasado con los *outcomes* de interés una vez que se ha implementado el programa. Por ejemplo, podemos llegar a saber, sin demasiadas dificultades, cuántos desempleados han encontrado trabajo después de participar en un curso de formación, cuántas patentes se han registrado en el marco de un programa de subvenciones de I+D+i o cómo han evolucionado las rentas de los agricultores después de un programa de apoyo a la tecnificación de un determinado tipo de cultivo.

Estimar Y_0, en cambio, es harina de otro costal. De hecho, construir un contrafactual apropiado es, de lejos, la tarea más complicada de la evaluación de impacto. El motivo de esta dificultad es, sencillamente, que el mundo no puede estar en dos estados al mismo tiempo: una ciudad no puede haber implementado un programa y no implementarlo al mismo tiempo, igual que una empresa no puede haber recibido una subvención de I+D+i y simultáneamente no haberla recibido. Si el programa se ha implementado, nunca podremos llegar a observar qué habría pasado si no se hubiera puesto en práctica. Por tanto, mientras que la estimación de Y_1 responde a una medida basada en la observación de la realidad, la estimación de Y_0 es siempre una declaración hipotética sobre cómo creemos que habría sido el mundo en ausencia del programa.

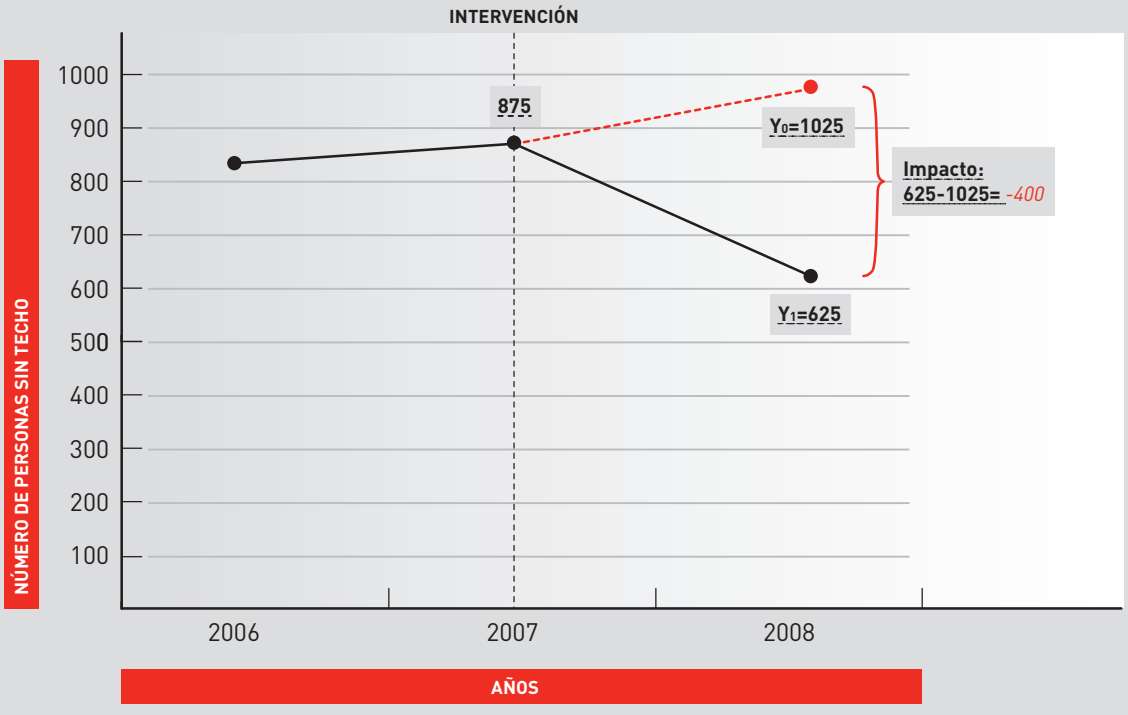
Así pues, la cuestión es: ¿cómo conseguimos formular una hipótesis contrafactual? El concepto en sí mismo no debería asustarnos, ya que la vida cotidiana está llena de ejemplos de este tipo de hipótesis: «Si hubiera estudiado más, habría aprobado las oposiciones»; o bien: «Si no me hubiera hipotecado, ahora no iría tan justo». El reto de la evaluación de políticas, sin embargo, es llegar a construir una hipótesis que no solamente parezca realista, sino que, además, permita cuantificar con precisión qué habría pasado en ausencia del programa, ya que necesitamos un número Y_0 con el que poder realizar la resta (Y_1 - Y_0) que nos lleva a estimar el impacto del programa.

Para hacerlo, la estrategia suele consistir en sustituir el contrafactual, que por definición no es observable, por un escenario de comparación observable. Por ejemplo, supongamos que el Departamento de Educación inicia un programa que consiste en otorgar autonomía de gestión a las direcciones de determinados centros escolares, con el fin de mejorar la calidad de la educación y, en último término, el rendimiento de los alumnos. Medir Y_1 es fácil: se trata de medir qué calificaciones han obtenido los niños de dichos centros escolares, un año después, por ejemplo, del cambio en el modelo de gestión. ¿Cuál puede ser la hipótesis contrafactual? Supongamos que en la red de centros escolares hay colegios de *características similares a los que han participado en el programa* que permanecen bajo el régimen de gestión ordinario. Podemos medir las calificaciones de los alumnos de esos centros similares y formular la siguiente hipótesis contrafactual: si los colegios que han participado en el programa no lo hubieran hecho (contrafactual no observable), las calificaciones que habrían obtenido sus alumnos serían las mismas que han obtenido los alumnos de los colegios de características similares que no han participado en él (escenario de comparación observable).

CUADRO 2
LA MEDIDA DEL IMPACTO CON UNA HIPÓTESIS CONTRAFACTUAL

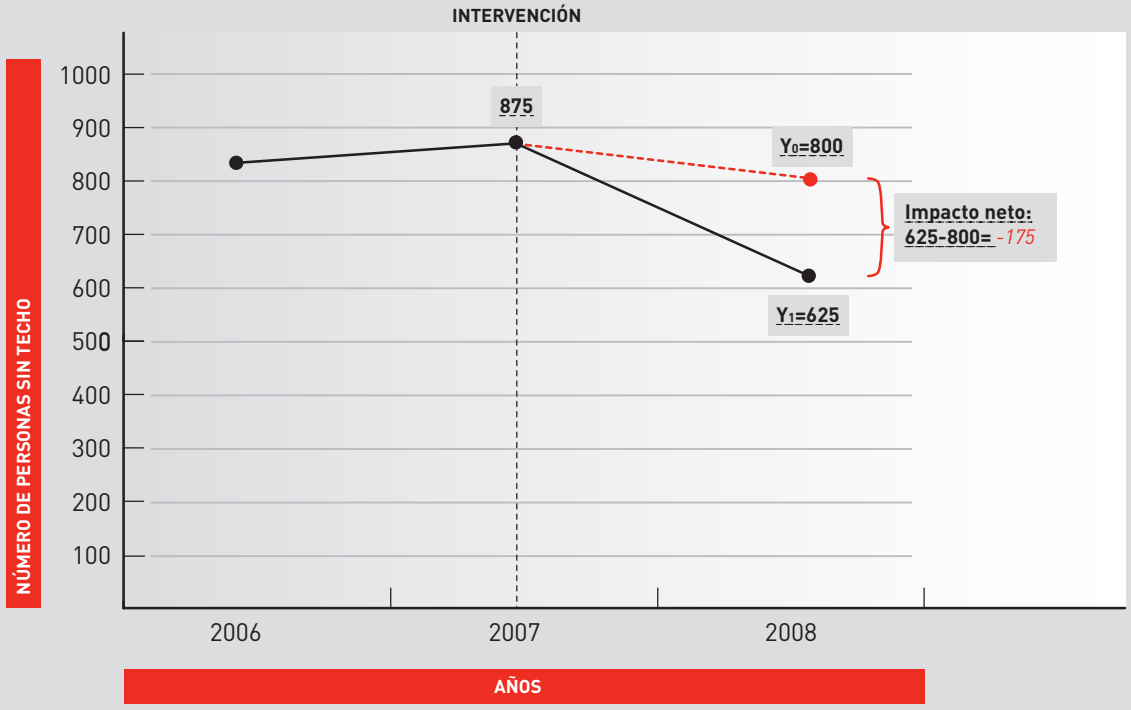
Los gráficos 2 y 3 —volviendo al ejemplo del programa de atención a las personas sin techo— representan con una línea roja dos posibles contrafactuales. El primero se basa en la estimación de que, en ausencia de la intervención, el número de personas sin techo habría aumentado (esto correspondería, por ejemplo, a un escenario de más desempleo, peores servicios a las personas con enfermedad mental y más inmigración indocumentada). Por otro lado, en la estimación del contrafactual del gráfico 3 se asume que el número de personas habría disminuido igualmente en ausencia del programa (a causa, por ejemplo, de un escenario de menos desempleo, mejores servicios y menos inmigración). Obsérvese que Y_1 no varía en ninguno de los dos gráficos: el programa comenzó, y luego se observó y midió el número de personas que pernoctaban en la calle después de la implementación. Por tanto, la divergencia de la magnitud del impacto en un gráfico y en el otro (400 personas en el gráfico 2 y 175 en el 3) se debe exclusivamente al hecho de que la estimación de Y_0 (el contrafactual) es diferente.

Gráfico 2. Evolución de las pernoctaciones en la calle con estimación del contrafactual (I)



Cuadro 2 (cont.)

Gráfico 3. Evolución de las pernoctaciones en la calle con estimación del contrafactual (II)



La bibliografía estadística y econométrica está repleta de estrategias para identificar el contrafactual de programas y políticas públicas, y en el capítulo 3 de esta guía expondremos las de uso más frecuente. Comprobaremos que el principal reto de estas **estrategias de identificación** radica en encontrar unidades (colegios, personas, barrios, etc.) que cumplan la condición de reunir *características similares* a las que han participado en el programa. Esto es debido a que, generalmente, si una persona participa en un programa y otra no, y si un barrio recibe una subvención y otro no, es porque son distintos en alguna característica relevante. Las estrategias de identificación del contrafactual hacen todo lo posible para controlar estas diferencias, con el inconveniente de que algunas son observables, pero otras no. Por ejemplo, podemos encontrar desempleados que se parezcan a los que han participado en un curso de formación en cuanto al nivel formativo previo, la historia laboral, la edad y otras características similares recogidas en una base de datos, pero no en lo que respecta a otros factores relevantes, como la motivación para encontrar trabajo, el estado anímico, etc.

Las metodologías para la evaluación de impactos que presentaremos en el capítulo 3 se adecúan a distintos tipos de programas y circunstancias de la evaluación, y no hay ninguna universalmente superior. La selección de la estrategia más adecuada requerirá, en cada caso, un análisis previo de las características de la intervención pública que la justifique, especialmente sobre los objetivos del programa, el procedimiento de selección de los participantes, el proceso de implementación y las fuentes de datos disponibles. Antes de exponer los distintos métodos para la evaluación de impacto, el capítulo 2 hace referencia a los pasos preliminares para enfocar el diseño de la evaluación, que guiarán la elección del método más adecuado.

CUADRO 3
LA ROBUSTEZ DE LAS HIPÓTESIS CONTRAFACTUALES

Las estrategias de identificación del contrafactual son hipótesis sobre situaciones que nunca se producirán, ya que, como hemos dicho, es imposible que una persona que ha participado en un programa al mismo tiempo no haya participado en él. Por tanto, todas las estrategias tienen en común que no pueden ser probadas empíricamente, es decir, nunca podremos comprobar *a posteriori* si eran correctas o falsas. Todo lo que podemos hacer es valorar si la hipótesis contrafactual parece más o menos realista y argumentar sobre los motivos por los que creemos que se trata (o no) de una hipótesis plausible. De hecho, las controversias sobre las evaluaciones giran casi siempre en torno a la robustez de la hipótesis contrafactual, es decir, sobre cómo de adecuado es el escenario de comparación identificado.

La bibliografía está llena de ejemplos de programas o políticas en los que distintas estrategias de identificación en la evaluación han conducido a estimaciones del impacto muy diferentes. Por ejemplo, las evaluaciones sobre la cooperación financiera internacional con los países en desarrollo han tendido a no detectar ningún impacto significativo sobre el crecimiento económico de los países receptores. Sin embargo, en el año 2000, los economistas del Banco Mundial Craig Burnside y David Dollar publicaron un artículo en el que introducían una novedad en este tipo de evaluaciones: la efectividad de las ayudas financieras podría depender de la calidad de las instituciones y las políticas fiscales, monetarias y comerciales del país receptor. Efectivamente, su evaluación indicaba que si la comparación se realizaba solamente entre países con una *buena gobernanza*, el impacto de la ayuda financiera era positivo y estadísticamente significativo. En cambio, en los países con instituciones y políticas deficientes, el impacto de la cooperación financiera era nulo. La influencia de esta evaluación fue muy importante, ya que varias instituciones empezaron a condicionar su cooperación financiera a la adopción, por parte de los países receptores, de las políticas y las instituciones identificadas como adecuadas en el artículo de Burnside y Dollar.

Otros artículos y evaluaciones posteriores han puesto en solfa la estrategia de identificación del contrafactual empleado en el mencionado artículo, con lo cual la pregunta de evaluación de fondo *(los países que reciben ayuda financiera internacional ¿se desarrollan económicamente más deprisa que si no la recibieran?)* sigue sin tener una respuesta clara.

2. PASOS PRELIMINARES PARA DISEÑAR UNA EVALUACIÓN DE IMPACTO

2.1. ¿ES OPORTUNO EVALUAR LOS IMPACTOS DEL PROGRAMA?

La evaluación de impacto es, en cierto modo, la reina de las evaluaciones. A pesar de la innegable importancia de realizar una evaluación de necesidades para caracterizar adecuadamente el problema que quiere abordarse, de ponderar bien el diseño de la intervención y asegurarse de que es robusto y coherente con el conocimiento que las ciencias sociales atesoran y de evaluar el proceso de implementación para detectar dificultades imprevistas y desviaciones respecto a las previsiones, pocos momentos son tan emocionantes, tanto para los gestores de los programas como para los evaluadores, como el de intentar responder a la pregunta: «*¿Funciona?*». Por eso puede existir la tentación de hacerse esta pregunta prematuramente, cuando las condiciones no son todavía adecuadas para poder realizar una evaluación de impacto, o cuando sería más aconsejable y relevante efectuar otro tipo de evaluación. Los requisitos para llevar a cabo una evaluación de impacto son los siguientes:

1. **El programa debe ser estable.** Para poder evaluar los impactos de una intervención pública, es muy conveniente que esta intervención se haya mantenido sin demasiadas variaciones durante cierto tiempo, ya que, de otro modo, resultará difícil determinar sobre cuál de las múltiples versiones del programa deberán estimarse los impactos. Además, en programas inestables o *volátiles*, es muy posible que los resultados de la evaluación de impacto resulten irrelevantes desde el mismo momento en que se conozcan porque la versión evaluada no coincide con la que se está implementando en aquel momento. La estabilidad del programa suele ser más baja cuando el programa es relativamente nuevo, ya que en las primeras fases de la implementación es habitual que se produzca cierto proceso de ajuste del programa por ensayo y error. En estas circunstancias, una evaluación de implementación que permita analizar de forma sistemática qué está pasando y detectar qué correcciones son precisas suele ser más útil que una evaluación de impacto. Al margen de esto, la excepción a esta regla la constituyen los programas piloto, que si bien por definición son siempre nuevos, se mantienen estables y fieles a su diseño original precisamente porque su objetivo es evaluar la efectividad.

2. **Es necesario haber descrito una teoría del cambio coherente.** Como cualquier otro tipo de evaluación, la evaluación de impacto requiere que previamente se hayan identificado los objetivos genuinos de la intervención (de lo contrario, no será posible determinar cuáles son los impactos que deben estimarse) y una teoría del cambio que enlace, de forma mínimamente plausible, las actividades y los productos del programa con los impactos que pretenden lograrse (ya que si esperar impactos positivos se demuestra poco realista, es preferible trabajar para mejorar el diseño de la intervención que para evaluar improbables

impactos). En otras palabras, antes de la evaluación de impacto (o en el marco de la evaluación de impacto) es necesaria una mínima evaluación del diseño.

3. **Es necesario tener un conocimiento adecuado del proceso de implementación.** El interés por saber si un programa funciona o no suele ir acompañado del interés por saber por qué funciona, motivo por el cual las evaluaciones de impacto a menudo se realizan junto con evaluaciones de la implementación. Pero incluso si nuestro interés se centra estrictamente en medir los impactos de la intervención, es necesario un mínimo conocimiento del proceso de implementación para interpretar los resultados de una evaluación de impacto y transformarlos en recomendaciones de mejora. Así, si una evaluación de impacto concluye que un programa no tiene ningún impacto significativo, es posible afirmar que la teoría del impacto que une los *outputs* con los *outcomes* era equivocada (véase Ivàlua, Guía práctica, 3[1]), o que el programa nunca llegó a implementarse como estaba planificado y los *outputs* previstos nunca llegaron a generarse, ya sea por desviaciones respecto al diseño o porque era imposible llevar a la práctica la teoría del proceso. Incluso si los resultados de la evaluación de impacto son positivos, comprobar que el proceso de implementación se ha producido de acuerdo con las previsiones refuerza la conclusión de que el programa es la causa de los impactos.

4. **Los impactos deben haberse podido producir.** Son raras las intervenciones públicas que producen impactos inmediatos, por lo que es necesario que transcurra cierto tiempo desde la implementación de la intervención antes de poder detectar el impacto. En las páginas que siguen veremos que una de las decisiones a tomar a la hora de diseñar una evaluación de impacto es escoger el momento más adecuado para medir el impacto, ya que es posible que algunos efectos tarden en producirse, tiendan a acumularse o desaparezcan con el tiempo. Si, dado el tipo de intervención, sabemos de antemano que este momento no ha llegado todavía, será preferible posponer la evaluación y esperar a que los impactos hayan podido producirse.

2.2. ¿A QUÉ NOS REFERIMOS CUANDO HABLAMOS DE *OUTCOMES*?

A lo largo de esta serie de guías metodológicas, hemos repetido en varias ocasiones que las políticas públicas tienen su razón de ser en la existencia de un problema o situación social insatisfactoria, y que los objetivos de la política pública deben hacer referencia al cambio que la intervención pública pretende inducir sobre este problema o situación. Parece, pues, que la definición de los *outcomes* con los que mediremos el impacto debería derivarse de forma bastante directa de los objetivos del programa, ya sean los declarados formalmente o los identificados en la elaboración de la teoría del cambio de la intervención. Por ejemplo, si el objetivo de un programa es la reducción de la *siniestralidad en las carreteras*, parece que la definición de los *outcomes* debería capturar de la mejor manera posible el fenómeno de la

siniestralidad en las carreteras. Sin embargo, la tarea de identificar los *outcomes* y la forma de medirlos raramente es directa y suele precisar de la toma de algunas decisiones sobre qué, cómo y cuándo medir.

En primer lugar, hay que tener presente que algunas intervenciones públicas tienen objetivos múltiples. Por ejemplo, la reducción de la velocidad máxima en los accesos a Barcelona tiene por objetivo reducir la contaminación y reducir los accidentes; y el Programa Interdepartamental de Rentas Mínimas de Inserción tiene por objetivo, como su propio nombre indica, elevar la renta de las personas beneficiarias de la prestación e insertarlas en el mercado laboral. Si este es el caso del programa que debemos evaluar, es necesario seleccionar el objetivo cuyos impactos nos interesa evaluar, o si decidimos evaluar más de uno, ser conscientes al planificar la evaluación de que esto supondrá una multiplicación de los recursos necesarios (tiempo, financiación, etc.).

Por otra parte, algunos objetivos son multidimensionales. Incluso si la intervención tiene un único objetivo, o si hemos elegido solo uno sobre el que queremos realizar la evaluación de impacto, las maneras de llegar a definir este impacto suelen ser múltiples. Supongamos, por ejemplo, que queremos capturar el fenómeno de la siniestralidad en las carreteras: podemos medir el número de accidentes, el de accidentes con heridos o muertos, o directamente el número de heridos y muertos en accidente de tráfico. Por el contrario, si queremos capturar el fenómeno de la inserción laboral, que suele ser el objetivo de las políticas activas de ocupación, las opciones se multiplican: nos puede interesar si la persona encuentra un trabajo dentro de un período de tiempo, o intentar capturar la conservación del puesto de trabajo, es decir, medir si la persona mantiene el trabajo al cabo de un tiempo determinado, o cuántos días en total ha trabajado a lo largo de este período de tiempo. Igualmente, es posible que nuestro interés en la inserción laboral sea instrumental, por lo que la dimensión que realmente nos resulta relevante es la variación en la renta o el incremento en el bienestar subjetivo derivados de la inserción laboral. En los términos que empleábamos en la Guía 3 sobre evaluación de diseño, la consecución de algunos objetivos implicaba lograr una secuencia previa de impactos (por ejemplo, encontrar trabajo, conservarlo, hecho que incrementa la renta y, en última instancia, el bienestar), que denominábamos *estructura de impactos*. Antes de iniciar la evaluación es preciso decidir cuál (o cuáles) de las múltiples dimensiones que constituyen esta estructura es la más relevante para el propósito de nuestra evaluación.

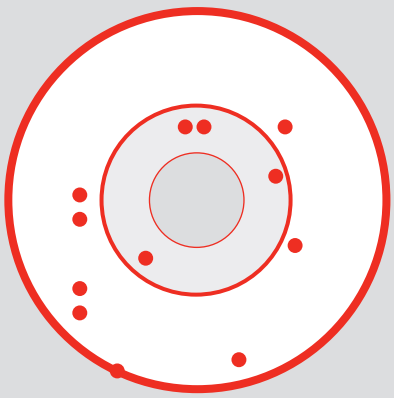
CUADRO 4
¡MEDIDAS!

«Supongamos que os contrato para medir un elefante. Puede parecer que la tarea está clara, pero, pensadlo un minuto. ¿Tenéis que medir su peso? ¿La altura? ¿Su longitud? ¿El volumen? ¿La intensidad de su color gris? ¿La cantidad y profundidad de sus arrugas? ¿O tal vez la proporción del día que se pasa durmiendo? Para poder medir a esta criatura necesitáis seleccionar una o unas cuantas características entre varias posibilidades. La elección dependerá de vuestro propósito a la hora de medir, o más bien del mío, ya que soy yo quien os ha contratado. Si yo fuera el responsable del transporte ferroviario de mercancías necesitaría conocer la altura, la longitud y el peso del elefante. Pero si fuera un taxidermista, estaría más interesado en su volumen y en las arrugas. Como domador, me preocuparía más qué proporción del día está dormido. Como productor de pieles sintéticas de animales, me gustaría saber el tono exacto del gris. Vosotros, viendo la oportunidad de mantener nos en nómina, seguramente insistiríais en el hecho de que no puedo entender a mi elefante si no conozco la variación estacional de la temperatura corporal.»

STONE, D. Policy Paradox, The Art of Political Decisión Making, 2002 [Traducción propia]

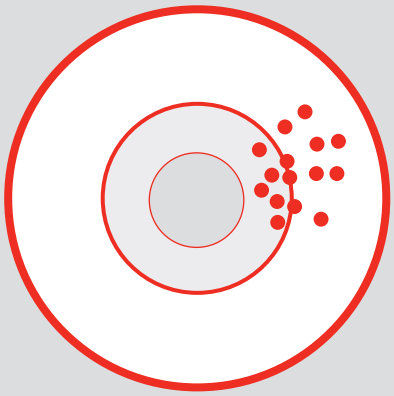
Por el contrario, algunos impactos son especialmente difíciles de medir porque los objetivos hacen referencia a constructos particularmente intangibles como, por ejemplo, incrementar la autonomía personal de los participantes de un programa de atención a las personas sin techo. En este caso, la dificultad no es tanto seleccionar una dimensión entre las varias que constituyen un objetivo, sino llegar a medir un fenómeno que, por su naturaleza, parece inmensurable. En estas situaciones, la decisión radica entre escoger una medida preexistente (existe a este respecto una bibliografía especializada en el desarrollo de medidas para los fenómenos sociales más diversos, desde el desarrollo cognitivo al estrés laboral, pasando por la felicidad y la percepción de seguridad en la vía pública) o crear una medida nueva ajustada a las especificidades de nuestra evaluación. En general, suele ser preferible escoger una medida preexistente, porque implica que alguien ya ha comprobado su **fiabilidad** (es decir, que si la medida se utiliza en diversas ocasiones, los resultados son coherentes), y también porque el uso de una medida estandarizada facilita la posterior comparación de resultados con otras evaluaciones. Además, el esfuerzo que requiere localizar una medida **válida** para nuestra evaluación en la bibliografía (o sea, que capture satisfactoriamente nuestro fenómeno de interés) suele ser sustancialmente menor que el de desarrollar y hacer pruebas con cuestionarios para elaborar una propia.

CUADRO 5
LOS CONCEPTOS DE VALIDEZ Y FIABILIDAD DE LA MEDIDA DEL IMPACTO



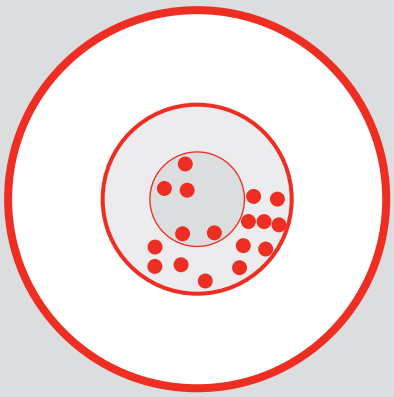
NI VÁLIDA NI FIABLE

La medida no captura el fenómeno de interés, y los distintos intentos de medir el fenómeno arrojan resultados dispares.



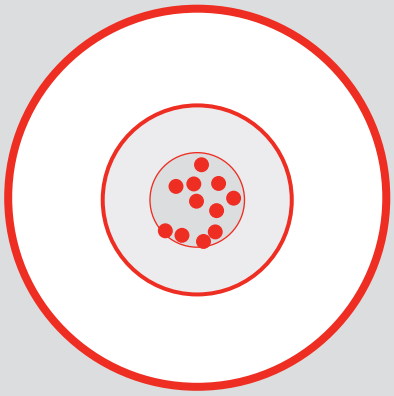
FIABLE, PERO NO VÁLIDA

La medida no captura el fenómeno de interés, pero los intentos repetidos de medir el fenómeno arrojan siempre resultados muy similares (pero equivocados).



REALMENTE VÁLIDA, PERO POCO FIABLE

La medida captura relativamente bien el fenómeno de interés, pero los distintos intentos de medir el fenómeno arrojan resultados demasiado dispares.



FIABLE Y VÁLIDA

La medida captura el fenómeno de interés, y los intentos repetidos de medir el fenómeno arrojan resultados muy similares.

Fuente: adaptación de VARKEVISSER, C.M.; PATHMANATHAN, I.; BROWNLEE, A. Designing and conducting health systems research projects. World Health Organization /International Development Research Centre, 2003.

Por último, debemos tener en cuenta que definir los impactos no solamente implica especificar qué medimos y cómo lo medimos, sino también cuándo lo medimos. Esta cuestión reviste una especial importancia, ya que distintos momentos de medida pueden conducir a conclusiones diferentes sobre los impactos del programa, ya que mientras que algunos impactos implican procesos lentos y pueden tardar en producirse, otros pueden ocurrir rápidamente, pero no mantenerse en el tiempo. En este sentido, se trata de determinar cuál es el momento más relevante para hacerse la pregunta: *«¿Qué ha pasado en comparación con lo que habría ocurrido si la intervención no se hubiera puesto en práctica?»*. Así pues, en un programa de ayuda a los funcionarios para dejar de fumar, el impacto puede ser fulgurante una semana después de comenzar, pero parece más relevante conocer el impacto un año después, ya que es probable que una parte de los que lo dejan inicialmente acaben recayendo. En cambio, una intervención para proteger el hábitat de una especie amenazada puede no presentar impactos apreciables en un principio, pero sí pueden ser muy notables al cabo de tres años, una vez que la población ha tenido tiempo suficiente para aumentar bajo las nuevas condiciones. En cualquier caso, el tiempo de medida deberá definirse con precisión: no se puede hablar de corto o largo plazo, sino que habrá que decidir, con exactitud, si nos referimos a seis, doce, dieciocho o veinticuatro meses después del programa.

2.3. ¿QUÉ QUIERE DECIR PARTICIPAR EN EL PROGRAMA?

Algunos conceptos y métodos de la evaluación cuantitativa de impactos han sido parcialmente importados de las ciencias médicas. Por ejemplo, en el ensayo de un medicamento se administra una píldora a algunas personas, que reciben el nombre de grupo de tratamiento o casos, y un placebo a las personas que constituyen el grupo de control, o simplemente *controles*. El efecto del medicamento se infiere a partir de la diferencia en la evolución de la patología o el síntoma de turno entre un grupo y otro. De forma análoga, los métodos para la evaluación de impacto suelen comparar un grupo de *tratamiento* (los colegios, personas, barrios, etc. que han participado en un programa) con un grupo de comparación o control (integrado por los que no han participado), que sirve para controlar el contrafactual.

Sin embargo, participar en un programa suele ser un concepto bastante más impreciso que tomarse una píldora. Mientras que con un medicamento no hay situaciones intermedias (sabemos si lo hemos administrado y en qué dosis), la participación en un programa puede querer decir cosas muy diferentes. Por ejemplo, que un barrio participe en la *ley de Barrios* quiere decir que ha recibido financiación para realizar determinadas actuaciones que son gestionadas de una manera determinada. La cantidad de financiación, el tipo de actuaciones y la forma en que se han gestionado cambia de un barrio a otro y, por este motivo, el que un barrio haya *participado* en la *ley de Barrios* se corresponde con situaciones muy diversas. Igualmente, que una persona haya participado en un curso de formación ocupacional puede querer decir desde que ha asistido al 100 % de las clases de un curso de jardinería de

80 horas, hasta que ha asistido al 50 % de las clases de un curso de mediación comunitaria de 20 horas. En síntesis, el *tratamiento*, en el caso de las políticas públicas, puede ser muy heterogéneo y plantea una cuestión a resolver: ¿de qué, exactamente, queremos estimar el impacto?

La **heterogeneidad del tratamiento** varía según el tipo de intervención pública. Si el nivel de variabilidad en lo que significa haber participado o haberse beneficiado de la intervención pública es importante, podemos emprender alguna de las siguientes medidas para tratarlo:

- Imponer restricciones a la definición de *participación* (por ejemplo, solamente considera remos que una persona ha participado en un curso de formación ocupacional si ha asistido como mínimo al 80 % de las clases de un curso de un mínimo de 30 horas).

- Desagregar la evaluación según el tipo de participación (por ejemplo, puede evaluarse el impacto de la formación en jardinería por separado del impacto de la formación en mediación comunitaria).

- Asumir la heterogeneidad del tratamiento como una característica de la intervención pública, teniendo siempre presente que se está infiriendo un impacto promedio de participaciones que en realidad son diversas.

2.4. ¿PARA QUIÉN QUEREMOS DETECTAR LOS IMPACTOS?

Muy a menudo, las intervenciones públicas se dirigen a una población diana bastante heterogénea. Por ejemplo, de la *ley de Barrios* pueden beneficiarse desde barrios de grandes zonas urbanas hasta otros de ciudades pequeñas, algunos cada vez más deshabitados y otros superpoblados. De forma similar, los programas de atención a las personas sin techo atienden desde personas con enfermedades mentales a inmigrantes recién llegados cuyo único problema es la falta de recursos y de una red social de apoyo, desde personas analfabetas a licenciados universitarios. Dada esta diversidad en la población diana, no es de extrañar que los programas puedan ser efectivos para determinados tipos de beneficiario, mientras que no lo son para otros. En este contexto de **heterogeneidad de los impactos**, estimar **impactos** promedios para todos los beneficiarios puede llevar a concluir que un programa es relativamente inefectivo para la mayoría de personas, cuando en realidad es muy efectivo para un subgrupo. En este caso, no se trataría tanto de descartar el programa, como de mantenerlo solamente para aquellos para los que resulta efectivo, y reformarlo para los demás.

Si los datos disponibles lo permiten, con la evaluación de impacto se podrá saber no solamente si el programa *funciona*, sino también para quién *funciona* mediante la desagregación de las estimaciones de impacto del programa para distintos subgrupos de población. Esto

permitiría averiguar, por ejemplo, si el carné de conducir por puntos resulta más efectivo a la hora de reducir la siniestralidad en las carreteras en el caso de los conductores jóvenes o de los de mediana edad, para infractores reincidentes o para los ocasionales, para los desplazamientos de ocio o los de trabajo.

Al preparar el diseño de una evaluación de impacto, es importante identificar cuáles son los subgrupos de población (por género, grupos de edad, tipo de problemática inicial, etc.) para los que resulta relevante realizar un análisis desagregado.

CUADRO 6
LAS DECISIONES METODOLÓGICAS EN EL PROCESO DE DISEÑO DE LA EVALUACIÓN

Diseñar una evaluación de impacto implica tomar decisiones constantemente: la definición del impacto, el momento de medida, la concreción de lo que significa participar en el programa, la desagregación del análisis por subgrupos o la elección del método para identificar el contrafactual no son pasos automáticos, sino que implican escoger una alternativa entre varias.

Cada una de estas decisiones conlleva resolver una disyuntiva. Por una parte, aumentar la complejidad del análisis (escoger más de una definición de impacto y momento de medida, desagregar el análisis en distintos grados de participación y subgrupos de beneficiarios, o evaluar el programa con más de una metodología) permite obtener información más detallada y extraer conclusiones más robustas. Por otra, incrementa el tiempo y los recursos necesarios para llevar a cabo la evaluación (a veces, hasta hacerla inabordable) y complica la comunicación de los resultados. Por consiguiente, incluso si decidimos que es asumible cierto grado de complejidad, resulta inevitable renunciar a algunas medidas del impacto, niveles de desagregación y aproximaciones metodológicas.

A pesar de que, idealmente, estas renuncias se asumen sobre criterios de menor relevancia, la toma de decisiones implica a veces cierto grado de arbitrariedad. Puede resultar difícil justificar por qué medimos la situación laboral al cabo de 12 meses y no de 24, por qué desagregamos el análisis por comarcas y no por grupos de edad, o por qué hemos escogido un método determinado en lugar de otro, y así hasta generar cierta sensación de que la fotografía que estamos ofreciendo sobre el rendimiento del programa es incompleta.

Al margen de las dudas que se planteen en la toma de estas decisiones metodológicas, lo más importante es tomarlas con diligencia, de modo que la evaluación esté acabada a tiempo para ser relevante, y hacer constar siempre bajo qué definición de impacto y en función de qué hipótesis de partida hemos llegado a la conclusión de que el programa es efectivo o no.

2.5. ¿A QUÉ NOS REFERIMOS EXACTAMENTE CUANDO HABLAMOS DE CONTRAFACTUAL?

Como hemos explicado anteriormente, la evaluación de impacto requiere identificar un escenario contrafactual con el que estimar los *outcomes* que se habrían producido en ausencia de la intervención pública. El concepto de contrafactual es, sin embargo, excesivamente ambiguo y requiere ser precisado antes de avanzar en el diseño metodológico de la evaluación:

- Por un lado, las intervenciones públicas muy pocas veces constituyen el primer intento de abordar un problema, sino que se trata de una reforma de un programa anterior. En este

contexto, el contrafactual es lo que habría ocurrido *si hubiéramos continuado con el programa antiguo.*

- A veces, sin embargo, el programa puede ser genuinamente nuevo, o puede interesarnos la estimación del impacto en relación con la ausencia de cualquier intervención pública. En estas situaciones, el contrafactual se convierte *en lo que habría pasado si no hubiera habido ningún programa en funcionamiento*.

- Finalmente, a veces, para un mismo objetivo, hay varios programas en funcionamiento, o hay uno que funciona con distintas variantes o modelos de implementación (por ejemplo, con provisión pública directa en unos lugares y externalizada en otros), por lo que el interés de la evaluación es valorar la efectividad de un programa o modelo respecto al resto. En estas situaciones, el contrafactual puede definirse en cualquiera de las dos versiones anteriores, dependiendo de la pregunta de evaluación y la aproximación metodológica para darle respuesta.

2.6. ¿DE QUÉ DATOS DISPONEMOS PARA HACER LA EVALUACIÓN DE IMPACTO?

La disponibilidad de datos determina, en la práctica, muchas de las decisiones sobre el diseño de una evaluación de impacto. No siempre podemos definir los *outcomes* como querríamos, sino que nos vemos forzados a definirlos de la mejor manera posible con los datos de los que disponemos. Lo mismo ocurre con la definición de lo que significa «participar en el programa», la identificación de los subgrupos de interés o la selección de la estrategia metodológica para controlar el contrafactual. La disponibilidad de datos es el mayor determinante de la tarea del evaluador, como el solar y el entorno lo son para el arquitecto.

Estas limitaciones se deben al hecho de que, en general, es preferible trabajar con datos preexistentes procedentes de registros administrativos. En efecto, si definimos los *outcomes*, los subgrupos y el tratamiento de manera que podamos extraerlos de datos preexistentes, la evaluación es mucho más rápida y barata que si tenemos que realizar una encuesta para generar datos nuevos. Además, si utilizamos datos administrativos, la muestra con la que trabajaremos será mucho más grande que si hacemos una encuesta, de modo que las estimaciones serán mucho más precisas. Por otra parte, nos ahorraremos los sesgos de no respuesta que sufren las encuestas y que complican el tratamiento estadístico de los datos (Purdon, 2002). No obstante, realizar una encuesta no siempre es una opción a descartar. A veces, para estimar el impacto de un programa necesitamos saber qué ha ocurrido con los participantes un tiempo después de que hayan abandonado el programa, cuando ya no se hace un seguimiento en los registros. La disyuntiva está siempre entre el coste, el tiempo y las limitaciones que implica generar datos nuevos mediante una encuesta, y la ventaja de poder recoger toda la información que nos interesa y de la forma que más nos interesa.

Sin embargo, las limitaciones que la calidad y el contenido de los registros administrativos imponen sobre las evaluaciones no deben ser consideradas como un designio inmutable. De acuerdo con un reconocido economista y evaluador del Banco Mundial, que las evaluaciones de impacto sean *ex post* por definición no quiere decir «que deban comenzar después de que el programa finalice, o ni siquiera después de que haya comenzado: las mejores evaluaciones *ex post* se diseñan y comienzan a implementarse *ex ante*» (Ravallion, 2006). Entre las medidas más importantes a tomar *ex ante* está la de conseguir que los registros administrativos incorporen información relevante para usos de evaluación y mejoren su calidad. Suele decirse que los problemas que no tienen solución no son problemas, sino condicionantes. En este sentido, la falta de datos adecuados para la evaluación en los registros administrativos es un condicionante a corto plazo y un problema a largo plazo.

Notas:

1 *BLASCO, J. Evaluación del diseño. Barcelona: Ivàlua, 2009. (Guías prácticas sobre evaluación de políticas públicas; 3)*

3. MÉTODOS PARA LA EVALUACIÓN DE IMPACTO

La cuestión fundamental que plantea la evaluación de impacto es medir hasta qué punto la aplicación de una determinada política sobre un conjunto de individuos modifica un determinado *outcome* de interés, como su renta o su salud, respecto de lo que estos mismos individuos habrían experimentado en ausencia de dicha política. Lo que complica la evaluación de impacto es que la situación en ausencia del programa, el denominado contrafactual, es algo que por definición resulta inobservable para el grupo de individuos que reciben el programa. Así pues, como ya hemos mencionado en el apartado anterior, el gran reto metodológico que plantea la evaluación de impacto es cómo definir a un grupo de individuos que, además de no participar o beneficiarse del programa o política, constituya un contrafactual creíble, es decir, que su nivel de *outcome* pueda considerarse equivalente al que habríamos observado para los beneficiarios de la política si esta no les hubiera sido aplicada.

Los métodos que se emplean en la evaluación de impacto difieren entre sí en función del procedimiento utilizado para definir el grupo de individuos que actúan como contrafactual:

- Por una parte, los denominados **diseños experimentales** son aquellos en los que, partiendo de una población de potenciales beneficiarios del programa o política, los individuos acaban participando o no de acuerdo con un mecanismo de asignación puramente aleatorio; los individuos que no participan, el denominado grupo de control, constituyen el contrafactual en este tipo de diseño.

- Por otra parte, el resto de métodos disponibles, que reciben el nombre de **diseños cuasiexperimentales**, comparten la característica de que la participación de los individuos en el programa no la define un procedimiento aleatorio: ya sea porque son los propios individuos los que eligen participar o no, ya sea porque otro agente toma esa decisión, o por las dos cosas al mismo tiempo. En los **diseños cuasiexperimentales**, el contrafactual se define a partir de los individuos que no participan en el programa, que constituyen lo que se denomina grupo de comparación.

Los apartados siguientes constituyen una introducción breve, de carácter no técnico, a los principales métodos que pueden utilizarse para establecer el impacto de una política[1]. Comenzaremos con una introducción de los dos principales retos que deben afrontar los distintos métodos: maximizar la robustez con que concluyen que el programa es la causa de los impactos observados (validez interna) y la potencialidad para generalizar las conclusiones a otros programas, situaciones y momentos (validez externa). A continuación, iniciaremos la exposición de los métodos con los experimentos sociales, ya que existe un amplio consenso en el sentido de que estos constituyen el diseño más robusto a la hora de evaluar el impacto de un programa. Por este motivo, y aunque son de uso poco habitual, representan el están-

dar que utilizan de espejo el resto de diseños. Los demás apartados explican los distintos métodos de carácter *cuasiexperimental* más utilizados: los diseños antes-después, la técnica de *matching* y el modelo de dobles diferencias.

3.1. LA VALIDEZ DE LAS CONCLUSIONES

3.1.1. LA VALIDEZ INTERNA

Los métodos para la evaluación de impacto que presentamos en este capítulo sirven para **inferir una relación causal** entre una intervención pública y determinados *outcomes*.

Utilizamos el concepto de **validez interna** para referirnos a la «verdad relativa» de una inferencia causal, es decir, a la robustez con que se concluye que el programa es el agente responsable de los impactos observados. La validez interna no es una propiedad de las metodologías, sino de las inferencias concretas que se realizan en cada evaluación, ya que un mismo método de evaluación puede producir conclusiones más o menos válidas según las circunstancias y características del programa evaluado.

Las **amenazas a la validez interna** son razones específicas por las cuales es posible que estemos parcial o totalmente equivocados a la hora de plantear una inferencia causal. Concretamente, son todas aquellas explicaciones alternativas, aparte del programa, que potencialmente podrían ser responsables de los cambios observados en los *outcomes*. En cada evaluación, diremos que el diseño metodológico es más o menos válido en tanto que descarte convincentemente estas explicaciones alternativas. Las listamos a continuación de forma separada, si bien algunas de ellas no son totalmente independientes:

1. **La historia / factores contemporáneos** Se refiere a todos los acontecimientos que suceden durante la implementación del programa y que pueden influir sobre los *outcomes*. En el ejemplo del programa de atención a las personas sin techo, las variaciones en el mercado de trabajo, la puesta en marcha de un programa de salud mental y los cambios en el control de la inmigración formaban parte de la historia del programa, ya que se producían al mismo tiempo que este e incidían sobre el número de personas que pernoctan en la calle (la medida del *outcome*), motivo por el cual podrían ser parcialmente responsables de los cambios observados y, por tanto, ser confundidos con el impacto del programa. Suele abordarse con la identificación de un grupo de comparación que esté expuesto a acontecimientos externos iguales o similares.

2. **El sesgo de selección.** Una cuestión crítica cuando se identifica a un grupo de comparación es que sea equivalente al grupo de participantes en todas las características que

están asociadas con los *outcomes*, excepto por el hecho de que unos participan en el programa y los otros no. El sesgo de selección se produce cuando esta situación no se cumple y existe, desde antes del programa, alguna diferencia significativa entre los participantes y el grupo de comparación que puede ser potencialmente responsable de las diferencias observadas al final del programa entre los *outcomes* de unos y otros. Imaginemos, por ejemplo, un programa de refuerzo lingüístico en catalán en el que se proporciona formación en lengua catalana solamente a los inmigrantes recién llegados que lo soliciten, con el objetivo final de facilitarles la inserción laboral. Es muy posible que los que se apunten sean distintos de los que no lo hagan en características relevantes para la inserción laboral: que su nivel educativo sea superior, que dominen mejor el castellano o que tengan más motivación para encontrar un trabajo. Es probable que, en ausencia del programa, los participantes tengan más fácilidades para acceder al mercado laboral que los no participantes. Por tanto, si comparamos la evolución de la participación en el mercado laboral de unos y otros es posible que parte de la diferencia en los *outcomes* se deba, en realidad, a estas diferencias iniciales en sus características. La amenaza del sesgo de selección es omnipresente en todos los diseños no experimentales, y abordarlo adecuadamente es, con diferencia, el principal reto metodológico de la evaluación de impacto.

3. **El desgaste diferencial de la muestra (*attrition*).** Se trata de una forma del sesgo de selección que se produce una vez iniciada la evaluación. Es relativamente habitual que, a lo largo de la evaluación, algunos participantes y miembros del grupo de comparación abandonen el programa, se nieguen a seguir respondiendo cuestionarios o simplemente desaparezcan. Estas pérdidas pueden llegar a cambiar la composición de los dos grupos de manera que es muy posible que acaben siendo distintos en alguna característica que esté relacionada con los *outcomes*, por mucho que inicialmente estuvieran equilibrados. Esta diferencia de composición entre un grupo y otro puede ser la responsable de los cambios observados en los *outcomes*, que, por tanto, pueden ser confundidos con el impacto del programa. Supongamos, por ejemplo, que en un programa destinado a prevenir recaídas en ex alcohólicos, los que mejor se encuentran y más seguros están de no recaer tienden a abandonarlo antes de su finalización porque lo consideran innecesario, y se les pierde la pista. En este caso, el grupo de participantes acaba estando compuesto por aquellos con un mayor riesgo de recaída, mientras que el grupo de comparación sigue constituido por una mezcla de personas con riesgos altos y bajos. En consecuencia, igual que ocurría con el sesgo de selección, es posible que parte de la diferencia en los *outcomes* entre los dos grupos se deba, en realidad, a estas diferencias finales en su composición.

4. **Regresión a la media.** Es la tendencia estadística que tienen los resultados extremos que se producen en un determinado momento de medida de los *outcomes* a acercarse a la media de la población cuando vuelven a ser medidos un tiempo después. Ello es así

porque muchos fenómenos implican cierta variación aleatoria: por ejemplo, a un fin de semana con muchos accidentes de tráfico suele sucederle otro con un número inferior, aunque las circunstancias que determinan la propensión a los accidentes (el clima, el volumen de tráfico, etc.) no hayan variado, del mismo modo que las personas que van a psicoterapia porque están muy estresadas es probable que la siguiente vez que vayan lo estén menos, aunque no hayan recibido tratamiento. En general, esta amenaza debe tenerse en cuenta si la selección para participar en el programa se produce precisamente porque la medida del *outcome* ha sido sustancialmente alta o baja. En estas situaciones, es muy probable que en la siguiente medida el *outcome* mejore por efecto de la regresión a la media, y que este efecto se confunda fácilmente con un efecto del programa.

5. **Efectos de los tests.** Algunas evaluaciones consisten en realizar un test a participantes y miembros del grupo de comparación antes del programa (pretest) y después (postest), con la finalidad de poder estimar cuál ha sido el impacto de la intervención. Ahora bien, hacer el pretest puede enseñar a las personas a hacerlo mejor en el test siguiente, o puede inducir otras formas de reacción que pueden confundirse con los impactos del programa. Por ejemplo, si el test consiste en hacer pruebas de colesterol, puede ser que las personas cuiden más su dieta porque saben que les volverán a medir. De la misma manera, en una prueba de vocabulario, es posible que las personas que han obtenido malos resultados se preparen para la siguiente porque les da vergüenza volver a hacerlo mal, o que sencillamente lo hagan mejor porque ya saben en qué consiste la prueba y tienen cierta práctica.

6. **El efecto Hawthorne.** Es un incremento del *outcome* que experimentan las personas por el mero hecho de que alguien les presta una atención especial, y no tanto por el efecto del programa en sí. Este efecto debe su nombre a una serie de estudios realizados entre los años 1927 y 1932 en los que se observó que los trabajadores de una planta eléctrica aumentaban su productividad cuando tenían la sensación de que la dirección se preocupaba por ellos, independientemente de la forma que tomara dicha atención. Así, tanto reducir la intensidad de la luz como subirla provocó los mismos impactos positivos.

7. **Maduración.** El cambio natural o el crecimiento debido al mero paso del tiempo pueden explicar las diferencias entre los *outcomes* medidos antes y después de un programa. Por ejemplo, la mejora de las capacidades cognitivas de los niños, el temperamento de los comportamientos de riesgo de los adolescentes o el empeoramiento de la autonomía personal de las personas mayores son fenómenos que se producirán entre el pretest y el postest por efecto de la maduración y que pueden confundirse con los impactos del programa. Para abordar esta amenaza es necesario disponer de un grupo de comparación de la misma edad para que el fenómeno de maduración afecte de manera similar a ambos grupos.

8. Efectos de los instrumentos Si se produce un cambio en el instrumento empleado para medir los *outcomes* en el pretest y en el postest, la variación en los *outcomes* puede reflejar los efectos de este cambio técnico en el sistema de recogida de datos y puede confundirse fácilmente con los impactos del programa. Es una amenaza frecuente cuando la evaluación hace un seguimiento de series temporales largas o cuando la medida depende de una valoración relativamente subjetiva que puede ir cambiando a lo largo del tiempo como, por ejemplo, la apreciación del grado de desestructuración de una persona sin techo en el momento de entrar al sistema.

9. Externalidades (*spillovers*). Se producen cuando los no participantes pueden absorber los beneficios del programa de forma indirecta, a menudo por el hecho de estar en contacto con los participantes. Por ejemplo, en un programa piloto de información sexual a adolescentes, es posible que el grupo de comparación mejore sus *outcomes* porque los participantes les han explicado lo que han aprendido. Esta amenaza lleva a la subestimación del impacto del programa.

3.1.2. LA VALIDEZ EXTERNA

La validez externa se refiere al grado en que las conclusiones de una evaluación pueden ser generalizadas a otros programas similares, momentos o lugares más allá de los propios de la misma evaluación. Por ejemplo, si una evaluación demuestra que una intervención para el fomento del espíritu emprendedor empresarial ha sido efectiva en un estado de tradición industrial, en los Estados Unidos, en el año 2006, ¿podemos concluir que también lo será ahora y en Cataluña?

Al igual que el diseño metodológico de la evaluación determina el grado de validez interna de las conclusiones, también lo hace con la validez externa. En general, cuanto más artificiales y controladas sean las condiciones del programa para facilitar la evaluación, menos plausible resulta pensar que estas condiciones se reproducirán en un programa similar que no esté sujeto a la evaluación, y menos generalizables serán las conclusiones. Así, la elección de un diseño metodológico deberá encontrar un equilibrio adecuado entre la validez interna y la externa.

3.2. EXPERIMENTOS SOCIALES

3.2.1. QUÉ SON Y QUÉ LOS HACE ROBUSTOS

Evaluar el impacto de una política pública mediante un experimento social es, desde una perspectiva metodológica, muy similar a aplicar la lógica que siguen los ensayos clínicos.

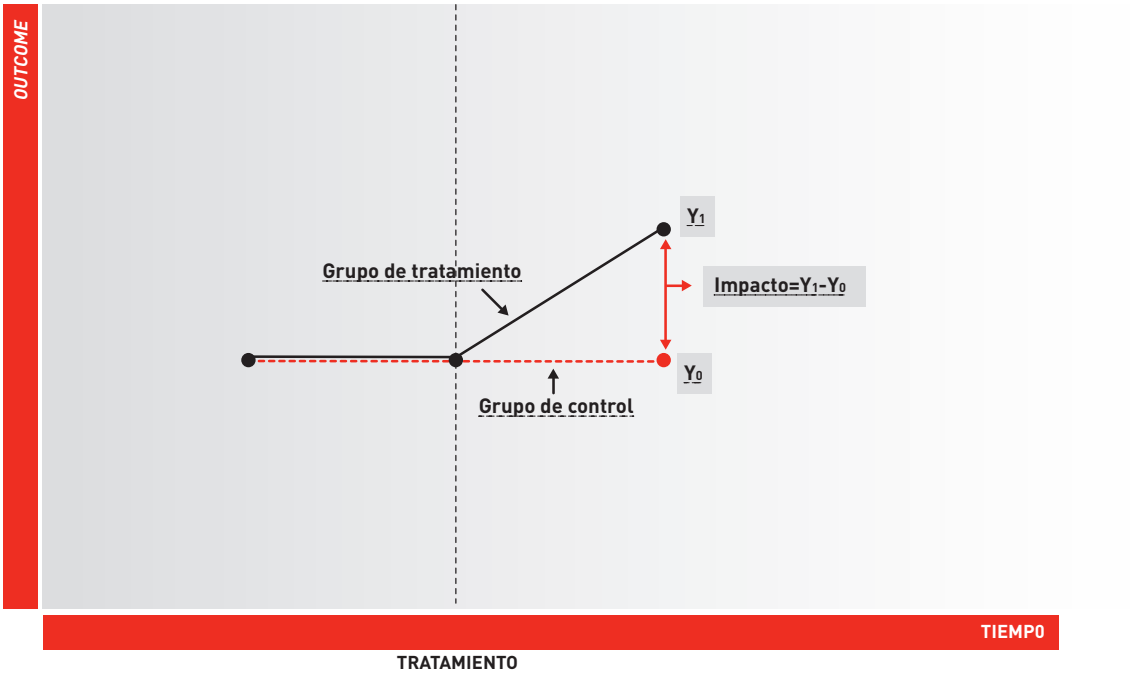
Así, después de seleccionar un conjunto de individuos susceptibles de beneficiarse de los potenciales efectos positivos de la política, se les asigna mediante un procedimiento aleatorio, con su consentimiento, a uno de los dos grupos siguientes: de un lado, el denominado grupo de tratamiento, en el que los sujetos participarán o recibirán durante cierto período de tiempo la intervención que caracteriza la política objeto de evaluación (un incentivo fiscal, un nuevo tipo de servicio, etc.); de otro lado, el denominado grupo de control, en el que los individuos no recibirán la intervención en cuestión. El hecho de no recibir la intervención no implica necesariamente que las personas que conforman el grupo de control no reciban ningún tipo de tratamiento: basta con que lo que reciban sea distinto de lo que prevé la política que estamos evaluando[2].

En este tipo de diseño experimental, el impacto de la política es muy fácil de medir: solamente hay que comparar, pasado cierto tiempo, la media que toma el *outcome* de interés (encontrar trabajo) entre los individuos que forman parte del grupo de tratamiento y los que integran el grupo de control (gráfico 4). Si esta diferencia de medias resulta estadísticamente significativa, podremos concluir que la política tiene un efecto (positivo o negativo) sobre el *outcome* que estemos analizando[3].

¿Qué es lo que explica que, a pesar de su sencillez, los experimentos sociales constituyan el diseño más robusto a la hora de medir el impacto de una política pública? El motivo hay que buscarlo en la asignación aleatoria, que consigue que los individuos del grupo de tratamiento y de control sean equivalentes en todos los factores que pueden influir sobre el *outcome* de interés, excepto en uno solamente: la participación o recepción de la política que estemos analizando. Además, el hecho de que algunos de estos factores sean inobservables para el analista o de difícil medición, como pueda ser la motivación o el interés de los individuos, resulta totalmente irrelevante: la aleatorización permite que los factores inobservables se distribuyan también de manera similar entre ambos grupos de personas. Así pues, dado que los dos grupos resultan equivalentes en todas aquellas variables (observables o no) que pueden influir sobre el *outcome* de interés, resulta legítimo atribuir *causalmente* cualquier diferencia en esta última variable a lo que distingue a los grupos entre sí: haber recibido o no la política.

En definitiva, la robustez de un experimento social como método para evaluar impactos deriva del hecho de que queda eliminada, por construcción, la principal amenaza a la validez interna de cualquier diseño de evaluación: la posible existencia de un sesgo de selección (véase el apartado 3.1.1). Pero esta gran virtud de los experimentos sociales solamente se acabará produciendo si la equivalencia entre el grupo de tratamiento y el de control se mantiene durante todo el período de tiempo que dura el experimento. En este sentido, como se explica en el siguiente apartado, existen diversas circunstancias que pueden aparecer durante la fase de implementación del experimento y que erosionan su validez, tanto interna como externa.

Gráfico 4. Ilustración de un experimento social.



Fuente: elaboración propia.

3.2.2. LÍMITES A LA VALIDEZ DE LOS EXPERIMENTOS

VALIDEZ INTERNA

- **Fracaso de la aleatorización.** La primera amenaza a la validez interna de un experimento social es que el proceso de asignación aleatoria de casos y controles no haya funcionado. La manera de comprobar este extremo pasa, simplemente, por contrastar si existen diferencias estadísticamente significativas entre ambos grupos en las medias de todas aquellas variables (observables) que pueden influir sobre el *outcome* (en el caso de un programa de inserción laboral, estas variables podrían ser la edad, el sexo, el nivel educativo, etc.). Es importante comprobarlo porque, en caso de que existan diferencias, podrían llevar a un sesgo de selección.

- **Sesgo de selección en las muestras (*Sample Selection Bias*).** Un problema con el que pueden encontrarse los experimentos sociales es que, a pesar de haber sido asignados aleatoriamente a los grupos de tratamiento y control, algunos de los individuos del primer grupo acaben no siguiendo el protocolo de tratamiento (por ejemplo, no asisten a los cursos de formación que prevé el programa), y/o algunos del grupo de control acaben teniendo acceso a dicho protocolo. El riesgo de que se produzca este tipo de situaciones depende de la naturaleza de la intervención que se esté analizando: así, aunque resulta plausible pensar que algunos de los tratados puedan decidir no asistir a cursos

de formación, parece poco probable que este rechazo se produzca si el tipo de intervención consiste en recibir una transferencia monetaria. Por otra parte, en cuanto a la posibilidad de que personas «control» acaben recibiendo la intervención, el aspecto clave a tener en cuenta es la capacidad de los responsables del experimento para monitorizar la actividad de los gestores del programa y evitar situaciones anómalas.

- **Externalidades (*spillovers*).** Cualquier efecto indirecto sobre los *outcomes* del grupo de control motivado por la existencia del tratamiento pone en entredicho la validez de los resultados generados por el experimento. Una selección precisa de las unidades a partir de las cuales se realizará el proceso de aleatorización puede prevenir este tipo de sesgo; a modo de ejemplo, si estamos interesados en medir el impacto de un programa escolar de salud alimentaria sobre la obesidad infantil, es evidente que la aleatorización no deberá realizarse entre individuos de un mismo colegio (habrá procesos de imitación), sino entre colegios que se encuentren a cierta distancia unos de otros.

- **Desgaste diferencial de la muestra.** En cualquier experimento social transcurre un lapso de tiempo entre el momento de la asignación aleatoria de los individuos a los grupos de tratamiento y control, y el momento en que se mide el *outcome* de interés para valorar el impacto de la política; si durante este lapso de tiempo hay individuos del grupo de tratamiento y/o de control que abandonan el experimento, de modo que resulta imposible medir sus *outcomes*, diremos que se ha producido un fenómeno de desgaste muestral. Este desgaste puede provocar un sesgo en la estimación del impacto si existen diferencias en las características de aquellos que abandonan respecto a los que permanecen, ya que desaparece la equivalencia entre los individuos del grupo de control y de tratamiento que se había conseguido en el momento de la aleatorización. En cualquier caso, para las situaciones en que se produce un desgaste muestral que puede amenazar la validez de los resultados, existen técnicas estadísticas que permiten corregir (parcialmente) el posible sesgo resultante.

La naturaleza prospectiva de los experimentos sociales hace que las fases de planificación y diseño de la evaluación sean de crucial importancia. El desgaste muestral, la existencia de externalidades y cualquier otro factor que pueda sesgar los resultados de la evaluación, deberán ser anticipados e incorporados al diseño del experimento para eliminarlos o minimizar su alcance. En caso contrario, cuando el experimento ya se encuentra en marcha, resulta prácticamente imposible rehacer el diseño.

VALIDEZ EXTERNA

En el caso de un diseño experimental, la validez externa de los resultados obtenidos (la posibilidad de extrapolarlos) puede verse afectada por dos motivos principales. En primer

lugar, puede ser que la muestra de individuos a partir de la cual se hayan definido los grupos de tratamiento y de control no sea representativa de la población a la que pretendemos extrapolar los resultados; este sería el caso, por ejemplo, de un experimento social que se hubiera llevado a cabo en una determinada comarca de Cataluña que no fuera representativa de la población catalana. Por otra parte, también puede pasar que el programa en sí mismo no resulte representativo, es decir, que la manera en que este opera en condiciones experimentales no pueda reproducirse a una escala superior (por ejemplo, en el caso de un programa de refuerzo educativo, puede ocurrir que el nivel de motivación de los profesionales no sea el mismo, o que la insuficiencia de recursos diluya algunos elementos del programa cuando se aplica a gran escala, etc.).

3.2.3. ¿POR QUÉ NO HAY MÁS EXPERIMENTOS SOCIALES?

Durante las últimas décadas, los experimentos sociales han experimentado un crecimiento notable, sobre todo en los países anglosajones y también en algunos países en vías de desarrollo. En los Estados Unidos, país abanderado en este ámbito, se han realizado experimentos sociales para evaluar cambios en las políticas educativas (Krueger, 1999), reformas en los programas de mantenimiento de rentas para personas pobres (Moffit, 2004), o también experiencias innovadoras en el campo de las políticas activas de ocupación (Heckman, 1997). También se han realizado experimentos sociales en otros países del continente americano, como por ejemplo en México, donde la evaluación experimental del programa PROGRESA tuvo un impacto notable a nivel internacional (Skoufias, 2005). En Colombia, Chile o Argentina, entre otros, también han evaluado mediante experimentos diversas políticas en ámbitos laborales, educativos o de los servicios sociales, así como en otros países del continente asiático y africano que reciben fondos procedentes de la ayuda internacional[4]. El cuadro 7 ilustra las características de este tipo de diseños de la mano de un experimento social concreto: la evaluación de una reforma organizativa de los servicios sociales y sanitarios para personas mayores en Quebec (Béland [et ál.], 2006).

En cualquier caso, dado que la mayoría de analistas considera los experimentos sociales el diseño más robusto para evaluar el impacto de una política (el *gold standard*), resulta hasta cierto punto paradójico que no existan muchos más experimentos sociales de los que hay.

Un primer factor a considerar es el elevado coste que, en general, tiene este tipo de diseños: por un lado, dado que la minimización de las amenazas a la validez del experimento requiere un riguroso proceso de planificación *ex ante*, la negociación entre las distintas partes implicadas sobre estas cuestiones puede resultar bastante importante en términos de tiempo; por otra parte, si la hipótesis es que los efectos de la política no sean inmediatos y se pretende poder extrapolar los resultados del experimento a otras áreas del país, habrá que trabajar con muestras de controles y tratamientos de dimensiones suficientes (miles de personas) que habrá que seguir durante un amplio período de tiempo[5].

En cualquier caso, más allá de las consideraciones económicas, el argumento habitual que utilizan los que se oponen a los experimentos sociales tiene un trasfondo ético: resulta inadecuado privar a determinados individuos (los del grupo de control) de los beneficios que supone una nueva política utilizando un mecanismo tan arbitrario como la aleatorización. La réplica por parte de aquellos que ven en los experimentos sociales una herramienta adecuada de evaluación se sustenta en tres consideraciones. La primera es que la presunción de que se está privando a algunos individuos de algo beneficioso no debería tener sentido si el experimento está justificado, ya que es precisamente la ausencia de datos sobre la efectividad del programa lo que justifica su evaluación. Por otra parte, son pocas las ocasiones en que pertenecer al grupo de control implica no recibir ningún tipo de intervención, sino que más bien lo que se compara es la nueva política respecto a «seguir como hasta ahora». Finalmente, hay situaciones bastante frecuentes en las que la aleatorización puede considerarse un criterio de asignación equitativo, como por ejemplo cuando la falta de recursos no permite atender de una sola vez a toda la población potencialmente beneficiaria de la política; de hecho, cuando se producen situaciones de este estilo, un diseño experimental más aceptable que utilizar una simple lotería entre individuos es optar por un despliegue aleatorizado (*randomized phase-in*): lo que se aleatoriza es el momento del tiempo en el que distintos grupos de individuos comenzarán a recibir el nuevo programa.

Al margen de cuáles sean las razones que subyacen tras la escasez de experimentos sociales, lo cierto es que muchas de las evaluaciones de impacto que se llevan a cabo en todo el mundo utilizan diseños de carácter no experimental. Dedicaremos los siguientes apartados a describir brevemente los principales métodos disponibles a tal efecto.

CUADRO 7
EJEMPLO DE EXPERIMENTO SOCIAL: SISTEMA INTEGRADO DE ATENCIÓN SANITARIA DE QUEBEC

CONTEXTO: en muchos casos, la falta de autonomía de las personas mayores viene motivada por el padecimiento de enfermedades crónicas y, por este motivo, las necesidades de atención de estos colectivos son tanto sanitarias como sociales. A pesar de ello, en la mayoría de países desarrollados —Canadá incluido— la respuesta asistencial que proporcionan los sistemas sanitario y social suele aplicarse sin ningún tipo de coordinación.

OBJETIVO: el equipo investigador pretendía evaluar en qué medida un sistema integrado de atención (SIPA, por sus siglas en francés) permitiría mejorar la salud de las personas mayores dependientes de Quebec, aumentar la satisfacción de sus cuidadores y reducir los costes asistenciales totales.

TIPO DE ESTUDIO E INTERVENCIÓN: la evaluación del nuevo modelo integrado de atención se llevó a cabo mediante un experimento aleatorio con grupo de control. Los pacientes asignados al grupo de tratamiento (606) pasaron a ser atendidos por equipos multidisciplinares que no solo proporcionaban directamente los servicios comunitarios sociales y sanitarios (atención domiciliaria, centro de día, centro de salud, enfermería domiciliaria, etc.), sino que también coordinaban la atención hospitalaria y la institucionalización social (residencias de asistidos) de los pacientes. Por otra parte, los individuos del *grupo de control* (624) continuaron recibiendo la atención de la forma habitual, o sea, mediante la acción independiente de los sistemas sanitario y social de Quebec.

***OUTCOMES*:** durante 22 meses, se obtuvo información de registro sobre los servicios sanitarios y sociales utilizados por los pacientes asignados a ambos grupos, incluyendo también los costes de la atención recibida en cada caso. Adicionalmente, en el momento de comenzar el estudio y transcurridos 12 meses, se utilizó una encuesta para obtener información sobre el estado de salud de la persona mayor, la satisfacción y la carga soportada por los cuidadores, así como sobre los gastos privados asumidos por la familia en relación con el cuidado de la persona dependiente.

RESULTADOS: los pacientes atendidos mediante el modelo SIPA utilizaron más los servicios sanitarios y sociales de carácter comunitario, pero su probabilidad de sufrir episodios de hospitalización innecesariamente largos (*bedblocking*) fue menor que la de las personas del grupo de control. Al margen de esto, en lo que respecta al resto de servicios sanitarios y sociales considerados, no se detectó ningún tipo de diferencia entre ambos grupos: utilizaron las urgencias hospitalarias con la misma intensidad, fueron ingresados en los hospitales con la misma frecuencia y tuvieron la misma probabilidad de acabar ingresados en una residencia de asistidos.

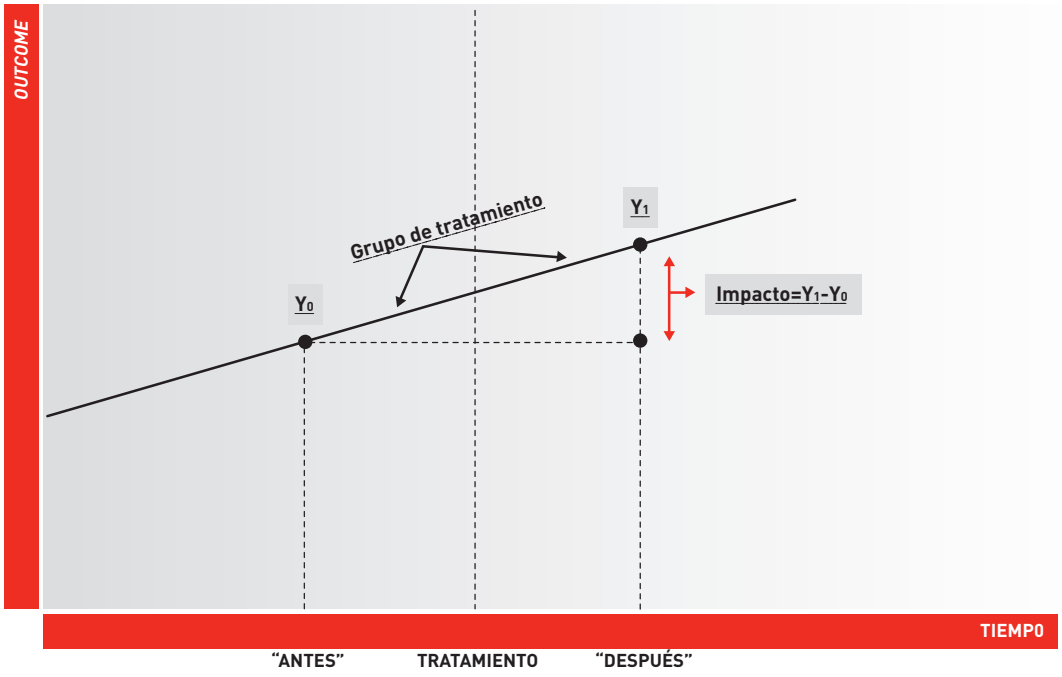
En términos de costes, si bien los pacientes del modelo SIPA tuvieron un gasto medio inferior en los servicios que implican la institucionalización de los individuos (hospitales y residencias), este efecto se vio totalmente compensado por un incremento en el gasto medio de los servicios comunitarios, de manera que el coste total medio de ambos grupos acabó siendo el mismo. Por otra parte, aunque la satisfacción de los cuidadores informales de los «pacientes SIPA» aumentó, no se detectaron diferencias significativas en cuanto a la «carga» soportada. Finalmente, tampoco se observaron diferencias entre ambos grupos en lo que atañe a la evolución del estado de salud de los pacientes tratados en cada caso.

Fuente: elaboración propia a partir de Béland [et ál.] (2006).

3.3. DISEÑOS SIN GRUPO DE CONTROL: ANTES-DESPUÉS Y SERIES TEMPORALES

El método cuasiexperimental más simple para evaluar impactos y, como veremos, también el menos robusto, es el denominado diseño antes-después. Su aplicación requiere disponer de información relativa a los beneficiarios de la política, tanto antes como después de su puesta en marcha. El impacto de la política se obtiene, simplemente, calculando la diferencia entre la media del *outcome* para la muestra de beneficiarios en cada uno de los dos momentos mencionados. El contrafactual se define reflexivamente, de aquí que este diseño se conozca también con el nombre de controles reflexivos, en el sentido de que la medida de «lo que les habría ocurrido a los beneficiarios en ausencia de la política» se obtiene a partir de la experiencia de estos mismos individuos antes de que la política existiera (gráfico 5).

Gráfico 5. Ilustración de un diseño antes-después.



Fuente: elaboración propia.

El supuesto clave para que este método estime correctamente el impacto de una política es que no puede haber ningún otro factor, salvo el programa, que haya podido afectar al *outcome* de interés entre los dos momentos de recogida de datos. En la mayoría de casos, sin embargo, resulta evidente que la plausibilidad de este supuesto será mínima. Imaginemos, a modo de ejemplo, una hipotética reforma que conceda más autonomía de gestión a los centros con el objetivo de reducir las tasas de fracaso escolar. En este caso, si nos aproximamos a la medida del impacto mediante un diseño antes-después, los posibles cambios que observamos en la evolución de las tasas de fracaso escolar pueden haber sido provocados por múltiples factores que no son la reforma: una reducción de las ratios alumnos/profesor fruto de la evolución demográfica, una reforma curricular, cambios en el perfil sociodemográfico de los padres, etc.

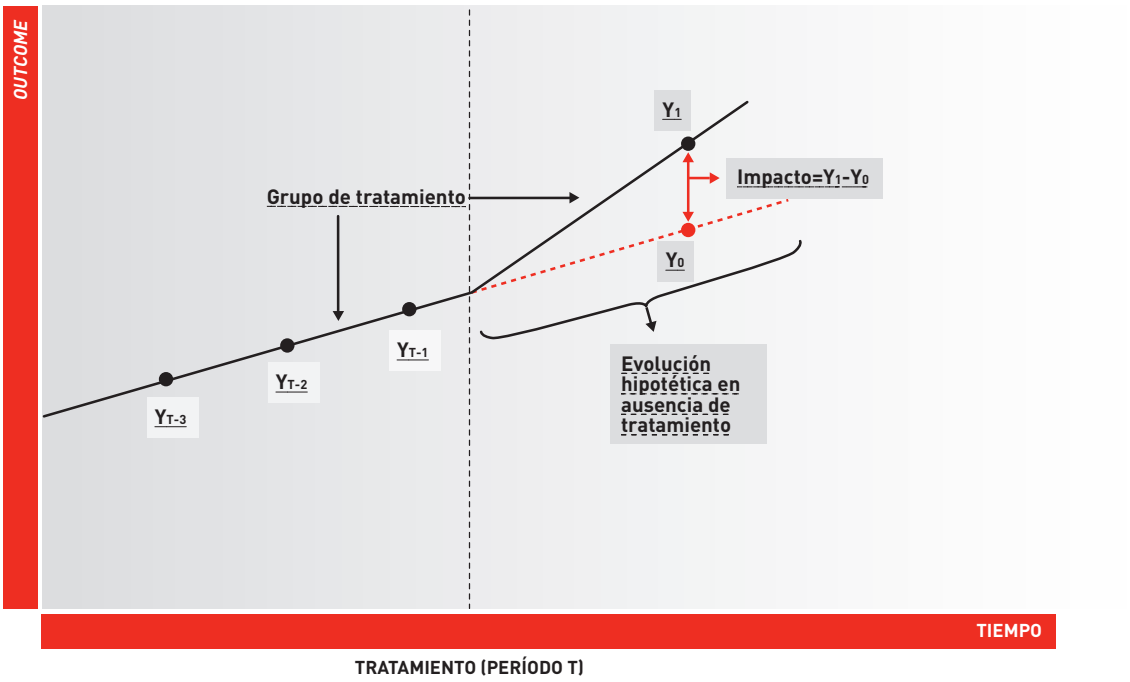
Pero, además de las amenazas a la validez interna provocadas por lo que en el apartado 3.1 denominábamos «historia o factores contemporáneos», los diseños antes-después son también muy vulnerables a las amenazas a la validez interna, especialmente los denominados fenómenos de maduración y de regresión a la media. En esencia, puesto que este tipo de diseño carece de un grupo de comparación genuino sobre el que construir un contrafactual creíble, siempre queda la duda de que las variaciones observadas en el *outcome* a lo largo del tiempo se habrían producido de todas formas, aunque la política evaluada no hubiera tenido lugar.

Así pues, a pesar de que se utilizan con bastante profusión, los diseños antes-después son un método muy poco robusto. Es por eso que, siempre que sea posible, optaremos por otros métodos que basen su estrategia de identificación en la comparación de grupos de personas beneficiarias y no beneficiarias de la política. ¿Qué hacer cuando resulta totalmente imposible construir un grupo de comparación no beneficiario de la política, como es típicamente el caso de una política que se introduce en todo el territorio y afecta a toda la población? En estas circunstancias, solamente si estamos muy seguros de que los impactos esperados de la política son bastante inmediatos y de que no hay factores contemporáneos que influyan sobre el *outcome*, podríamos llegar a considerar un diseño antes-después; en cambio, si estas circunstancias no se dan, habría que reconsiderar seriamente la conveniencia de llevar a cabo una evaluación de impacto cuantitativa.

Los denominados modelos de **series temporales interrumpidas** constituyen el otro gran tipo de diseño cuasiexperimental que, al igual que los diseños antes-después, intenta estimar el impacto de una política sin utilizar un grupo de comparación. En cierta medida, constituyen una variante refinada de los diseños antes-después, ya que su principal característica es que utilizan información sobre múltiples períodos de tiempo, tanto anteriores como posteriores a la introducción de la política que se pretende evaluar. Así pues, en comparación con un modelo antes-después, el contrafactual reflexivo de este tipo de diseños resulta más creíble, ya que disponemos de más información para estimar qué habría pasado en ausencia de la política.

La estrategia de identificación de los impactos que utilizan los diseños de series temporales interrumpidas es sencilla. A partir de las observaciones disponibles sobre la evolución del *outcome* antes de la intervención, se utilizan técnicas estadísticas para intentar modelizar su comportamiento «normal» en ausencia de la intervención, teniendo en cuenta la posible influencia que hayan podido tener otros factores. A continuación, este comportamiento normal se proyecta en los períodos posteriores a la introducción de la política, y se contrasta hasta qué punto existen discrepancias entre las predicciones del modelo y los valores realmente observados; si las hay, se atribuyen a la existencia de la política (gráfico 6). No obstante, aunque la idea subyacente es simple, hay que decir que los modelos de series temporales son técnicamente complejos y su aplicación exige conocimientos avanzados de estadística. d'estadística.

Gráfico 6. Ilustración de un diseño de series temporales interrumpidas.



Fuente: elaboración propia.

En cualquier caso, aunque los modelos de series temporales constituyen un método más robusto que los diseños antes-después para evaluar impactos, hay que tener presente que el contrafactual se sigue construyendo de forma reflexiva. Así pues, aunque podamos tener en cuenta la influencia de otros factores sobre la evolución del *outcome*, esta influencia se modeliza en función de una información que no pertenece al mismo período de tiempo en el que realmente opera la política. En definitiva, hay que seguir aplicando el mismo tipo de cautelas —si bien algo más suavizadas— que planteábamos en el caso de los diseños antes-después, es decir, limitar su uso a situaciones en las que, por un lado, no resulte posible construir un grupo de comparación y, por otro, haya una cantidad muy reducida de explicaciones alternativas sobre el porqué de la evolución del *outcome* después de la introducción de la política. En el resto de casos, resulta recomendable explorar otros tipos de diseños, como los que se explican a continuación, en los que el impacto se mide utilizando personas no beneficiarias (grupo de comparación) contemporáneas de las que se benefician de la política (grupo de tratamiento).

3.4. LA TÉCNICA DEL *MATCHING*

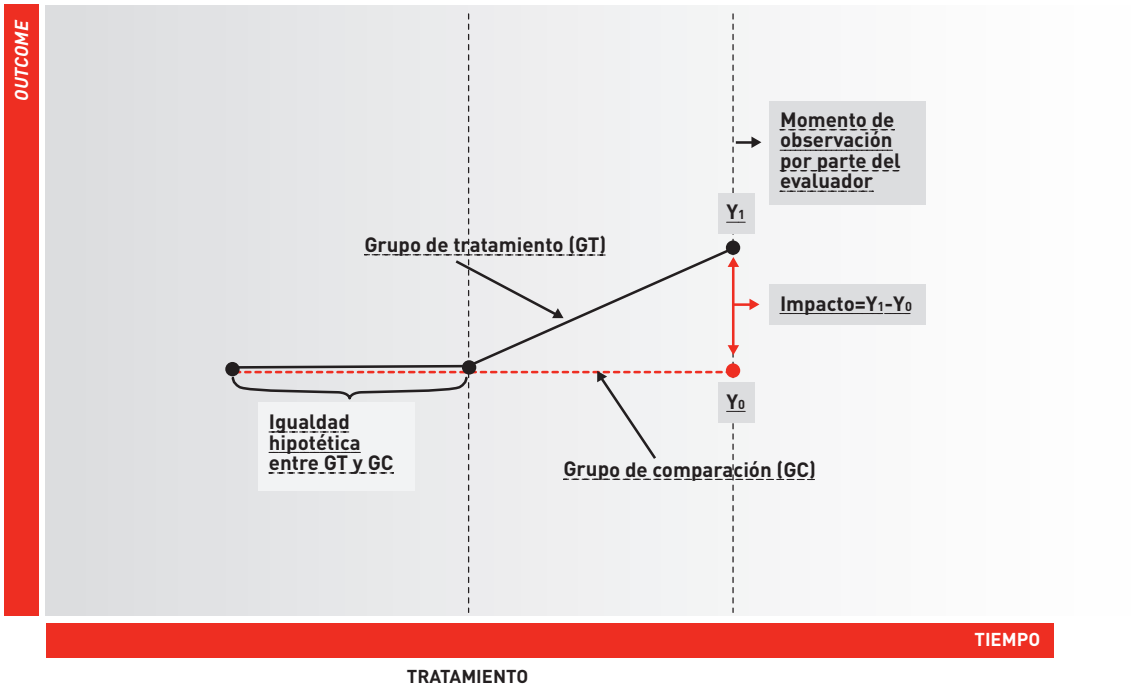
3.4.1. ¿QUÉ ES?

Esta técnica imita un experimento con asignación aleatoria de tratamiento mediante la creación de un grupo de control *ex post* que se parece lo máximo posible al grupo de tratamiento en cuanto a características relevantes observables. La aplicación de este método para evaluar el impacto de una política puede considerarse en aquellos casos en que, con posterioridad a la intervención pública, disponemos de información tanto de una muestra de individuos que han sido beneficiarios del programa como de otra de personas que no lo han sido. En concreto, para cada uno de los individuos de ambos grupos, hay que tener información sobre el valor que toma en cada caso el *outcome* de interés y también sobre todos aquellos factores (características de los individuos, entorno en el que viven, etc.) que, por un lado, pueden haber determinado el proceso por el que los individuos han decidido participar en el programa y, por otro lado, pueden tener efectos sobre el valor que toma el *outcome* de interés.

Lo que propone el método de *matching* es utilizar toda la información anterior para construir un grupo de comparación entre los individuos que no se benefician del programa. Para hacerlo, el método busca, para cada uno de los individuos que componen la muestra de tratados, una pareja o *match* (de aquí el nombre de la técnica) que sea lo más parecida posible en el sentido que acabamos de describir.

La pretensión que hay detrás de la técnica del *matching* es obtener, mediante procedimientos estadísticos, lo que los experimentos sociales obtienen mediante la asignación aleatoria, a saber, que el grupo de individuos que utilicemos para construir el contrafactual sea lo más parecido posible al grupo de individuos que reciben el programa, con el fin de minimizar tanto como se pueda el sesgo de selección. Pero mientras que una asignación aleatoria verdadera distribuye de forma *equitativa* las características observables y las no observables entre el grupo de control y el de tratamiento, el *matching* solamente distribuye *equitativamente* las características observables. En otras palabras, asume que no hay ninguna variable relevante no observable que difiera sistemáticamente entre el grupo de tratamiento y el de comparación y que, por tanto, el *outcome* del grupo de tratamiento si no hubiera participado o no se hubiera beneficiado del programa (es decir, el contrafactual) equivale al *outcome* del grupo de comparación que, realmente, no ha participado (gráfico 7).

Gráfico 7. Ilustración de un diseño basado en la técnica del *matching*.



Fuente: elaboración propia.

Un ejemplo puede ayudarnos a acabar de comprender la lógica de este tipo de diseño. Imaginemos que el Departamento de Salud pone en marcha una política de incentivos destinada a incrementar la prescripción de genéricos por parte de los médicos de atención primaria. Supongamos que sobre el porcentaje de medicamentos genéricos que prescriben los facultativos solamente influye la edad del médico y su sexo, y que la decisión de participar o no en el programa de incentivos es voluntaria. Dada esta situación, si tuviéramos la suerte de que no hubiera diferencias en lo que respecta al sexo y la edad de los médicos que participan y de los que no, una simple comparación de medias entre ambos grupos respecto al porcentaje de prescripción de genéricos nos proporcionaría una buena estimación del impacto del esquema de incentivos. ¿Y si, por el contrario, observamos que la distribución por sexo y edad de los participantes es distinta de la de los no participantes? Entonces no podríamos atribuir la diferencia en la media de los *outcomes* exclusivamente a la intervención, ya que estará también motivada por el hecho de que ambos colectivos son diferentes. En este caso, una posible estrategia sería construir el grupo de comparación seleccionando únicamente a aquellos médicos no participantes que garantizaran un porcentaje de mujeres y una distribución por edades idénticos a los del grupo de participantes: así, para cada mujer participante de entre 30 y 35 años, buscaríamos una mujer de igual edad en el grupo de no participantes. Una lógica muy similar a la que acabamos de describir es la que utiliza el *matching* para intentar obtener estimaciones no sesgadas del impacto de las políticas.

3.4.2. CÓMO SE IMPLEMENTA: *PROPENSITY SCORE* Y EMPAREJAMIENTO

El ejemplo anterior es poco realista en el sentido de que resulta evidente que la prescripción de genéricos se encuentra determinada por otros factores aparte del sexo y la edad de los médicos. En general, dado que el número de variables susceptibles de influir tanto sobre la decisión de participar en un programa como sobre el *outcome* de interés será bastante elevado, resulta imposible realizar un emparejamiento como el que describíamos en el ejemplo anterior.

La alternativa está en reducir la dimensionalidad del problema y definir la mayor o menor similitud entre tratamientos y controles a partir de un único número: el denominado *propensity score (PS)*. El PS mide la probabilidad de que un individuo, dadas sus características, decida participar en el programa; esta probabilidad se obtiene a partir de un modelo de elección discreta, como un logit o un probit[6].

El paso siguiente consiste en realizar los emparejamientos entre participantes y no participantes basándonos en el PS de unos y otros. Existen diversos métodos para definir cómo se constituyen las parejas. El más sencillo es el que se denomina el vecino más cercano *(nearest-neighbour caliper)* y consiste en formar tan solo aquellas parejas participante-no participante en las que la diferencia entre el PS de uno y otro sea inferior a cierto número predeterminado. Este método permite que estén equilibradas la muestra de participantes y la muestra final de no participantes con la que les acabamos comparando, por lo menos en lo que respecta a las variables que hemos considerado a la hora de modelizar la participación en el programa.

El último paso consiste en estimar el impacto de la intervención. En este sentido, igual que en el caso de los experimentos sociales, la técnica de cálculo es bien sencilla: basta con computar la media aritmética de las diferencias en *outcomes* de las distintas parejas construidas, y verificar si esta media es significativamente diferente de cero o no.

3.4.3. LIMITACIONES

El supuesto básico que la técnica del *matching* necesita para obtener estimaciones consistentes del impacto de una política es que, en media, una vez que se ha tenido en cuenta el efecto de las variables condicionantes (el sexo, la edad, la especialidad, etc., en el caso del ejemplo de los médicos), los participantes habrían obtenido el mismo *outcome* que los no participantes si la política no hubiera existido. O, dicho de otro modo, el supuesto fundamental es que no existe lo que técnicamente se denomina selección en variables no observables, es decir, no existe ningún factor que no haya sido tenido en cuenta por el analista que influya simultáneamente sobre la probabilidad de participar en el programa y

sobre el *outcome* de interés. En caso contrario, puesto que no hay nada que garantice que el emparejamiento haya generado muestras de tratamientos y controles equilibradas en lo que respecta a estos factores no observados, la medida del impacto que obtenemos puede sufrir un sesgo importante respecto a su auténtico valor. En este sentido, siguiendo con el ejemplo de los incentivos a los médicos, este sería el caso si existieran diferencias (no observables) de motivación entre participantes y no participantes.

Intuitivamente, para minimizar el riesgo de que se produzca un sesgo de selección en las propias estimaciones, parece obvio que lo que debería hacer el analista es intentar aplicar la técnica del *matching* utilizando un conjunto de variables de control lo más amplio posible; en concreto, deberían tenerse en cuenta todas aquellas variables de las que existiera evidencia de que influyen tanto sobre la participación como sobre el *outcome* de interés. En este sentido, si para algunos de estos factores no existe información (es decir, si estos factores son inobservables), la credibilidad de los resultados obtenidos quedará erosionada.

Con la intención de ilustrar las posibilidades que ofrece la técnica del *matching* en la práctica, el cuadro siguiente contiene la descripción de una evaluación de impacto que, siguiendo esta metodología, intentó averiguar la efectividad de los principales programas de formación ocupacional existentes en Cataluña.

CUADRO 8
EVALUACIÓN DE LA FORMACIÓN OCUPACIONAL EN CATALUÑA

El Servicio de Ocupación de Cataluña (SOC) desarrolla un amplio conjunto de acciones formativas dirigidas a diversos colectivos de desempleados, cuyo objetivo es mejorar las posibilidades de que estas personas encuentren un trabajo y lo mantengan. Los programas en marcha comprenden, entre otros, los siguientes: *Plan FIP*, destinado prioritariamente a personas desocupadas mayores de 65 años, desempleados de larga duración, discapacitados, etc.; *Centros de Innovación y Formación Ocupacional (CIFO)*, especializado cada uno de ellos en una o varias familias profesionales; *Igualdad de Oportunidades*, programa de formación dirigido específicamente a mujeres; etcétera. En el año 2008, por encargo del SOC, un equipo de investigadores dirigido por el profesor Toharia realizó una evaluación de los impactos de estos programas utilizando la técnica del *matching* (Toharia [et ál.], 2008). Los principales ingredientes metodológicos de esta evaluación fueron los siguientes:

- ***Outcomes***: situación laboral de la persona durante los ocho trimestres posteriores al año en que tuvieron lugar los programas evaluados.

- **Grupos de tratamiento y de control:** se definieron ocho grupos de tratamiento distintos, uno para cada uno de los ocho programas de formación ocupacional evaluados (Plan FIP, CIFO, Igualdad de Oportunidades, etc.). Adicionalmente, se definieron mediante la técnica del *matching* ocho grupos de comparación constituidos por demandantes de ocupación que no se habían beneficiado de ninguno de los programas formativos del SOC, pero que según su *propensity score*, presentaban características similares a las personas beneficiarias de los distintos programas.

- **Variables del *propensity score*:** sexo, edad, nacionalidad, nivel de estudios, ámbito de investigación, tiempo de inscripción, alta nueva, desempleados de larga duración, recepción de prestaciones, número de ocupaciones demandadas y provincia de residencia.

Los resultados obtenidos indican que tanto los CIFO como el Plan FIP aumentan la probabilidad de estar ocupados de los participantes respecto a los no participantes. En cambio, en lo que respecta al programa de Igualdad de Oportunidades y a las Acciones Integradas (dirigidas a personas con dificultades especiales), los impactos estimados sobre la ocupación fueron nulos. Finalmente, en el caso de los Programas de Garantía Social, dirigidos a jóvenes que finalizan la ESO sin acreditarla, se detectó un efecto negativo sobre la probabilidad de estar ocupado que tiende, sin embargo, a disminuir rápidamente en el tiempo; ahora bien, debe tenerse en cuenta que este programa es el de mayor duración y, por tanto, hay que pensar que los efectos tienden a producirse a más largo plazo.

Fuente: elaboración propia a partir de Toharia [et ál.] (2008).

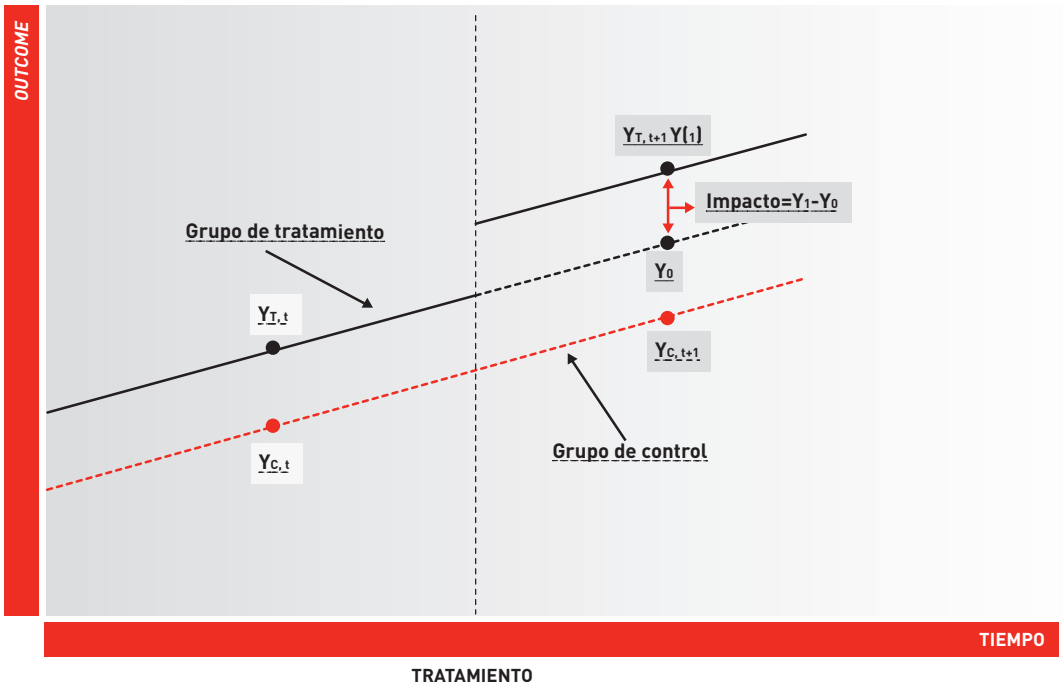
3.5. EL MODELO DE DOBLES DIFERENCIAS

3.5.1. DEFINICIÓN Y VENTAJAS

Esta técnica se aproxima a la cuantificación del impacto de un programa definiendo el efecto no en términos de la diferencia postratamiento en el nivel del *outcome* para los be-

neficiarios y para los no beneficiarios, sino como la diferencia en la variación del *outcome* antes y después de la política en ambos grupos[7]. Así pues, al definir el impacto de esta manera, la técnica de dobles diferencias reconoce explícitamente que parte de la variación temporal en el outcome de los que reciben la política se habría producido en cualquier caso, y que la manera de medirla es a través del cambio en el *outcome* de los no beneficiarios durante el mismo período. La mejor forma de entender esta técnica es a través de su representación gráfica.

Gráfico 8. Ilustración de un modelo de dobles diferencias.



Fuente: elaboración propia.

Así, tal como puede observarse en el gráfico 8, el impacto que se estima con un modelo de dobles diferencias es la diferencia entre el *outcome* de los beneficiarios (tratamientos) después de la política ($Y_{T,\ t+1}$; nuestro Y_1) y el valor de este *outcome* para este mismo colectivo en ausencia del programa, el famoso contrafactual, representado en el gráfico como Y_0. La esencia del método es que este contrafactual se obtiene proyectando el nivel del *outcome* de los beneficiarios antes de la política ($Y_{T,\ t}$) en una determinada tasa de variación: la observada en lo que respecta a los *outcomes* de los «controles» entre el momento anterior ($Y_{C,\ t}$) y posterior a la introducción de la política ($Y_{C,\ t+1}$). En definitiva, es fácil demostrar que la medida del impacto puede expresarse en términos analíticos como la «diferencia de diferencias» (de aquí el nombre de la técnica) siguiente:

$$(Y_{T,\ t+1} - Y_{T,\ t}) - (Y_{C,\ t+1} - Y_{C,\ t})$$

donde $Y_{T,\ t}$ y $Y_{T,\ t+1}$ son las medias del *outcome* para el grupo de tratamiento antes y después de la política, y $Y_{C,\ t}$ y $Y_{C,\ t+1}$ las del grupo de comparación.

El modelo de dobles diferencias (DD), en la medida en que utiliza información de antes y de después de la puesta en marcha de la política, tanto para los beneficiarios como para los no beneficiarios, es capaz de superar algunas de las limitaciones que amenazaban la validez interna de otros tipos de diseños[8].

En primer lugar, si lo comparamos con un diseño antes-después, el hecho de que el modelo DD utilice un grupo de comparación permite prevenir el posible sesgo provocado por factores contemporáneos a la política que pueden tener efectos sobre el *outcome* de interés. Por ejemplo, si estamos interesados en evaluar el efecto de un programa de formación sobre las posibilidades de encontrar trabajo de los desempleados, factores de este tipo serían las variaciones en la tasa de desempleo, modificaciones en la normativa laboral, etc. Igualmente, fruto de la existencia de un grupo de comparación, un modelo de DD que analizase el impacto de un programa de becas sobre el rendimiento escolar sería también menos sensible que un diseño antes-después a sufrir sesgos por regresión a la media (ya que a este fenómeno afectaría igual a beneficiarios y no beneficiarios).

Por otra parte, en la medida en que lo que se estima es la diferencia entre tratamientos y controles en la variación de los *outcomes* y no la diferencia en el nivel en sí, los modelos DD pueden eliminar algunas de las fuentes del sesgo de selección que la existencia de factores inobservables provocaba en el caso del *matching*. En concreto, el tipo de factores inobservables que no tienen efecto sobre la consistencia de la medida de impacto de un modelo DD son aquellos que no varían a lo largo del tiempo. Podemos ilustrar esta propiedad a partir del ejemplo mencionado sobre un hipotético programa de formación. Supongamos que la motivación fuera un factor inobservable y que esta variable se distribuyese de manera distinta entre los individuos que participan en el programa (más motivados) y los que no lo hacen (menos motivados). En este caso, es evidente que parte de la diferencia en el nivel de los *outcomes* de ambos grupos (tanto antes como después de la intervención) se explicaría por la influencia de este factor; ahora bien, como el impacto que mide el modelo DD no se da en términos de niveles, sino de tasas de variación del *outcome*, el hecho de que la diferencia de motivación no varíe a lo largo del tiempo hace que este factor no pueda haber sido la causa de la evolución diferencial del *outcome* en el grupo de tratamiento respecto al de control.

3.5.2. LIMITACIONES

Los modelos de dobles diferencias, a pesar de sus ventajas, no se encuentran exentos de ver amenazada su validez interna si no se cumplen los dos supuestos que permiten a este tipo de diseño identificar correctamente el impacto de una política pública.

El primero de estos supuestos es que tanto los participantes como los no participantes deben reaccionar de la misma manera ante los factores contemporáneos a la política que,

más allá de esta, pueden influir sobre el *outcome* de interés. En el caso del programa de formación antes mencionado esto significa que, por ejemplo, si se produce una mejora en un factor que influye sobre la probabilidad que tienen los individuos de encontrar trabajo, como pueda ser una reducción en la tasa de desempleo, su efecto sobre tratamientos y controles deberá ser del mismo. En este caso, la violación de este supuesto podría producirse si el aumento de la ocupación se hubiera concentrado en trabajos de elevada cualificación, y los niveles formativos de los tratamientos fueran superiores a los de los controles, ya que entonces la mejora inducida por la caída del desempleo sería superior entre los primeros.

Existen dos formas de intentar mitigar las posibles sospechas que puedan existir sobre el cumplimiento del supuesto «de igualdad de reacción ante factores contemporáneos». En primer lugar, si existe información sobre múltiples períodos de tiempo previos a la introducción de la política, podemos contrastar si efectivamente los *outcomes* de tratamientos y controles han evolucionado de manera similar cuando se han producido variaciones en determinados factores que también tienen influencia sobre el *outcome* (la tasa de desempleo, en nuestro ejemplo). La otra posibilidad que podemos aplicar cuando no existe información retrospectiva es estimar el modelo DD después de haber seleccionado los grupos de tratamiento y control utilizando la técnica del *matching*. De esta manera, dado que el *matching* garantiza una elevada similitud entre los dos grupos, hay que pensar que las posibilidades de que unos y otros reaccionen de la misma manera ante factores contemporáneos aumentan y, por tanto, también la consistencia de los resultados del modelo DD.

El segundo supuesto que deberá satisfacerse para que el modelo DD proporcione estimaciones no sesgadas del impacto de una política es que no pueden existir diferencias entre tratamientos y controles en características no observables que varíen a lo largo del tiempo. Si existen, el hecho de que los modelos DD miden el impacto como la diferencia entre tratamientos y controles en la variación del *outcome*, no permite en este caso eliminar posibles sesgos. Así pues, siguiendo con el ejemplo del programa de formación, si la motivación de tratamientos y controles varía a lo largo del tiempo, y no podemos observar esta variable, no podremos estar plenamente seguros de que este factor no es la causa de la evolución diferencial del *outcome* en el grupo de tratamiento respecto al de control y, por tanto, de que la magnitud del impacto estimado para la política no sobrestime su efecto real. Por consiguiente, si queremos que los resultados de una evaluación de impacto que utilice un diseño DD resulten creíbles, tendremos que presentar argumentos que permitan descartar la existencia de características inobservables que varíen en el tiempo de forma distinta entre tratamientos y controles.

El cuadro siguiente ilustra las posibilidades de los modelos DD a través de una aplicación llevada a cabo en nuestro entorno. En concreto, el caso comentado es el de una evaluación de impacto que estima, mediante un modelo DD, los efectos que podrían derivarse de una mayor cobertura por parte del sector público de la atención bucodental de los niños.

CUADRO 9
EVALUACIÓN DEL PROGRAMA DE ATENCIÓN DENTAL DEL PAÍS VASCO

El Programa de Atención Dental Infantil (PADI) del País Vasco, que lleva en funcionamiento desde el año 1990, constituye una experiencia de referencia en España, ya que ofrece un nivel de cobertura pública en lo que respecta a la atención dental muy superior a la que se observa en el resto del Estado. Este programa, además de cubrir las extracciones como en el resto de comunidades autónomas, incluye también una revisión anual y el tratamiento de caries y malformaciones en todos los niños del País Vasco de entre 7 y 15 años.

García (2005) realizó una evaluación del PADI que pretendía averiguar los efectos de este programa sobre los tres *outcomes* siguientes: la probabilidad de no haber ido nunca al dentista, de haber ido en los últimos tres meses y, finalmente, que la última visita fuera una revisión. El estudio estima el impacto del programa sobre estas variables utilizando un modelo de dobles diferencias. En concreto, partiendo de dos ediciones de la Encuesta Nacional de Salud correspondientes a los años 1987 y 2001, la autora obtiene información anterior y posterior a la introducción de la política tanto para el grupo de tratamiento (los niños del País Vasco) como para el grupo de comparación (los niños del resto de comunidades autónomas). Los resultados obtenidos sugieren que el programa solamente ha mejorado uno de los tres *outcomes* considerados: la probabilidad de haber ido al dentista en los últimos tres meses, superior en el grupo de tratamiento (niños del País Vasco) respecto al de control (niños del resto de comunidades autónomas).

Fuente: elaboración propia a partir de García (2005).

3.6. ELECCIÓN ENTRE MÉTODOS

Los apartados anteriores han puesto de manifiesto la existencia de varios métodos susceptibles de ser utilizados a la hora de intentar establecer el impacto de una determinada política. En general, una visión bastante compartida entre los evaluadores es que no existe el método ideal, es decir, un tipo de diseño en particular que, independientemente de las circunstancias, debería aplicarse de forma universal en todas las evaluaciones de impacto (Rossi, Lipsey y Freeman, 2004). En la práctica, por tanto, los evaluadores se ven obligados a decidir entre varias alternativas. Un elemento obvio que condiciona estas elecciones es la disponibilidad de tiempo y de recursos, pero hay otros: las características del programa, la importancia y el uso que se espere hacer de los resultados, la disponibilidad de datos, etc. Los apartados que siguen tratan brevemente sobre estos aspectos, y argumentan a favor de la necesidad de aproximarse a la elección del método con una mentalidad abierta, ecléctica y desprovista de excesivos apriorismos.

3.6.1. CARACTERÍSTICAS DEL PROGRAMA Y DISPONIBILIDAD DE DATOS

Hay determinadas características de las políticas públicas que aumentan las posibilidades de medir su impacto con rigor. Una especialmente importante es la relativa a su novedad y, más concretamente, a su concepción como prueba piloto. En estos casos, si se reúnen

una serie de condiciones, como que la demanda potencial sea superior a la oferta o existan dudas sobre la efectividad del programa, los experimentos sociales que utilizan procedimientos de asignación aleatorios pueden constituir una forma de evaluación de impacto a considerar. En cualquier caso, a pesar de que la asignación no se produce de manera aleatoria, un programa piloto que se implante solamente en determinadas zonas geográficas abre las puertas a diseños no experimentales (*matching* o modelos DD) que utilicen las áreas no piloto para construir grupos de comparación.

De todas formas, incluso en aquellos casos en que una nueva política se implementa sin pruebas piloto y afecta de repente a todo el territorio, siguen existiendo posibilidades de construir grupos de comparación si, por los motivos que sea, no toda la población potencialmente beneficiaria acaba participando en el programa. El peor de los casos se produce, desde la perspectiva de la evaluación de impacto, cuando una nueva política se implanta a escala nacional y afecta a toda la población, ya que esto solamente permite la aplicación de métodos reflexivos (antes-después y series temporales).

Otra ventaja de las políticas nuevas, se materialicen o no mediante pruebas piloto, es que permiten la introducción de elementos de evaluabilidad mientras se desarrolla la fase de diseño del programa. Como hemos mencionado anteriormente, una evaluación de impacto es, por definición, una evaluación *ex post*, pero las mejores evaluaciones de impacto son aquellas que se planifican *ex ante*. La posibilidad más extrema es que el mismo despliegue de la política se realice pensando en la evaluación, como es el caso de un experimento social, pero a veces basta con planificar una buena recogida de datos antes y después de la intervención, que afecte a sendas muestras de potenciales beneficiarios y no beneficiarios, para incrementar enormemente las posibilidades de obtener estimaciones de impacto creíbles mediante métodos no experimentales.

Pero a menudo el impacto que se desea evaluar no es el de una política nueva. En estos casos, dado que resulta imposible influir en «clave evaluadora» sobre el diseño del programa, el reto de la evaluación consiste en encontrar características de la política y fuentes de información que hagan posible la aplicación de las técnicas cuasiexperimentales descritas en esta guía.

Así pues, en lo que respecta a las características del programa, hay que buscar elementos que posibiliten la construcción de contrafactuales: por ejemplo, si por los motivos que sea un determinado programa tiene listas de espera, los individuos incluidos en ella pueden constituir un grupo de control natural respecto del que estimar el impacto del programa; asimismo, en la medida en que exista variabilidad geográfica en el grado de implantación de una política, las unidades territoriales que dispongan del programa pueden compararse con las que no lo tienen (las comunidades autónomas pueden constituir, en el caso de algunas políticas, una fuente de variabilidad a explorar en este sentido).

Por otra parte, respecto a la disponibilidad de fuentes de información, la impresión general que se tiene en nuestro país es que existe una infrautilización de los registros administrativos con finalidades evaluadoras. En este sentido, una vez que se tiene claro el diseño que puede tomar la evaluación de la política o programa, la tarea del equipo evaluador consiste en identificar todas aquellas bases de datos con información relevante sobre los individuos que componen los grupos de control y tratamiento previamente definidos, idealmente con el horizonte temporal más amplio posible. Igualmente, además de los registros administrativos, la búsqueda de información puede ampliarse a encuestas ya disponibles o, incluso, a la elaboración de una nueva.

3.6.2. ECLECTICISMO

Existen bastantes casos en los que el equipo evaluador, una vez exploradas las características del programa y las fuentes de datos disponibles, se dará cuenta de que pueden utilizarse varias de las técnicas cuasiexperimentales comentadas en los apartados previos, y no solamente una. En estas cuasiexperimentales, excepto los diseños que no utilizan grupos de comparación, poco recomendables como ya se ha comentado, no existe evidencia concluyente de que haya una determinada metodología que domine claramente sobre el resto[9]. Es por este motivo que, en general, los evaluadores acostumbran a aplicar simultáneamente varios tipos de metodologías, solución que permite verificar adicionalmente hasta qué punto los resultados obtenidos dependen mucho o poco de las elecciones de carácter metodológico.

Las distintas técnicas en que hemos centrado nuestra atención hasta el momento son metodologías de análisis cuantitativas. No es extraño que este tipo de enfoque sea preeminente en la evaluación de impacto, ya que la cuestión fundamental a resolver, que no es otra que la construcción de un contrafactual, es de naturaleza básicamente cuantitativa. A pesar de esto, existe la percepción creciente entre los evaluadores de que, con el fin de mejorar la robustez de la evaluación de impacto, resulta recomendable complementar el análisis utilizando técnicas cualitativas (entrevistas en profundidad o grupos de discusión). El valor añadido que puede aportar su utilización es permitir al equipo evaluador mejorar su conocimiento sobre las condiciones en que realmente opera el programa, las perspectivas de sus beneficiarios y otros elementos fundamentales a la hora de entender realmente el porqué del impacto de una política o programa (o de su ausencia).

Notas:

1 *El lector interesado puede profundizar en el estudio de estos métodos siguiendo las lecturas recomendadas que aparecen en el anexo de esta guía. También encontrará referencias que tratan sobre dos técnicas que, dado su carácter más técnico, hemos optado por dejar fuera de una guía de carácter introductorio: los modelos con variables instrumentales y el diseño de regresión discontinua.*

2 *A lo largo de la exposición, nos referiremos de manera genérica a individuos «tratados» y «controles», a pesar de que en muchas situaciones las unidades de análisis no son personas. Es lo que ocurriría, por ejemplo, si quisiéramos evaluar una política de incentivos fiscales destinados a empresas para aumentar su investigación en I+D+i, o una reforma que diera más autonomía de gestión a los centros escolares.*

3 *No entraremos en los detalles relativos a la dimensión (número de personas) que deben tener las muestras que componen los grupos de control y tratamiento, ya que se trata de una cuestión muy técnica. Solamente mencionaremos que cuanto más grande sea el tamaño de estas muestras, más posibilidades habrán de detectar la existencia de efectos atribuibles a la política, por muy pequeños que estos sean. Para una discusión detallada de estas cuestiones, véase Purdon (2002).*

4 *Un listado muy amplio de evaluaciones de impacto hechas en todo el mundo, tanto con diseños experimentales como cuasiexperimentales, puede encontrarse en la página web del Banco Mundial que aparece referenciada en el anexo de esta guía.*

5 *El elevado coste de un experimento no constituye, por sí solo, un argumento suficiente para decidir no llevarlo a cabo. La comparación relevante debe realizarse teniendo en cuenta también las consecuencias que puede suponer ampliar una política que, a pesar de no tener ningún impacto demostrado, absorbe una cantidad ingente de recursos públicos.*

6 *Los modelos de elección discreta son aquellos que pretenden establecer la relación existente entre una variable dependiente binaria (por ejemplo, participar o no) y una serie de variables independientes que* a priori *se considera que pueden influir sobre aquella. La diferencia entre los dos modelos mencionados radica en la forma funcional que se supone que relaciona la variable dependiente con las independientes: una función logística en el caso del logit, una función normal del caso del probit. Para obtener más detalles sobre este tipo de modelos, véase Corbetta (2007).*

7 *Este tipo de modelos se conoce en inglés con el nombre de* difference-in-differences*, aunque a menudo se utiliza la abreviatura diff-in-diff para referirse a ellos. Hemos optado por traducirlos por modelos de dobles diferencias siguiendo la propuesta de traducción al castellano sugerida por Vera-Hernández (2003).*

8 *Es importante señalar que, a la hora de estimar impactos mediante un modelo DD, no es necesario que la información sea longitudinal (es decir, sobre los mismos individuos antes y después de la intervención). Pueden utilizarse datos de sección cruzada (dos encuestas realizadas antes y después de la intervención a individuos diferentes), siempre y cuando podamos identificar a beneficiarios y no beneficiarios en uno y otro momento.*

[9] *La manera en que se evalúa la robustez de los métodos de evaluación de impacto cuasiexperimentales es aplicándolos a bases de datos que han sido obtenidas a partir de un experimento social. Así pues, partiendo de la premisa de que el experimento social permite identificar el impacto real, los resultados obtenidos para el resto de métodos se comparan con estos.*

BIBLIOGRAFÍA

BÉLAND, F. [et ál.]. «A system of integrated care for older persons with disabilities in Canada: Results from a randomized controlled trial». *The Journals of Gerontology: Medical Sciences* (2006), n.º 61 (4), pp. 367-373.

CORBETTA, P. *Metodología y Técnicas de Investigación Social*. Madrid: MacGrawHill, 2007.

GARCIA, P. «Evaluación de un Programa de Atención Dental Público: PADI en el País Vasco». *Ekonomiaz* (2005), n.º 60, pp. 62-89.

HECKMAN, J.; HIDEHIKO, I.; TODD, P. «Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme». *Review of Economic Studies* (1997), n.º 64 (4), pp. 605-654.

KUEGER, A. B. «Experimental Estimates of Education Production Functions». *The Quarterly Journal of Economics* (1999), n.º 114, pp. 497-532.

MOFFITT, R. A. «The Role of Randomized Field Trials in Social Science Research. A Perspective from Evaluations of Reforms of Social Welfare Programs». *American Behavioral Scientist* (2004), n.º 47 (5), pp. 506-540.

PURDON, S. *Estimating the impact of labour market programmes*. Londres: Department for Work and Pensions, 2002. (Working Paper n°. 3)

RAVALLION, M. *Evaluating Anti-Poverty Programs*. Washington DC: World Bank, 2006. (Policy Research Working Paper 3625)

SKOUFIAS, E. *PROGRESA and Its Impact on the Welfare of Rural Households in Mexico*. Washington DC: International Food Research Institute, 2005. (Research Report 139)

TOHARIA, L. [et ál.]. *Estudio de evaluación de la formación ocupacional en Catalunya*. Barcelona: Servei d'Ocupació de Catalunya, 2008. (mimeo)

ANEXO. GUÍA DE RECURSOS

MANUALES

MANUALES ESPECÍFICOS DE EVALUACIÓN DE IMPACTO:

BAKER, J. *Evaluating the Impact of Development Projects on Poverty—A Handbook for Practitioners*. Washington, DC: World Bank, 2000.

ASIAN DEVELOPMENT BANK. *Impact Evaluation: Methodological and Operational Issues*. Manila: ADB, 2006.
(http://www.adb.org/Documents/Handbooks/Impact-Analysis/default.asp)

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company, 2002.

MANUALES GENERALES DE EVALUACIÓN CON CAPÍTULOS SOBRE EVALUACIÓN DE IMPACTO:

ROSSI, P. H.; LIPSEY, M. W; FREEMAN, H. E. *Evaluation: a systematic approach*.
7.ª ed. Londres: Sage, 2004.

WEISS, C. *Evaluation*. 2.ª ed. Nueva Jersey: Prentice Hall, 1998.

ARTÍCULOS

La mayoría de artículos que se mencionan a continuación, y otros relacionados, pueden descargarse gratuitamente desde la siguiente página web del Banco Mundial:

http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20215333~menuPK:451260~pagePK:148956~piPK:216618~theSitePK:384329,00.html

ARTÍCULOS INTRODUCTORIOS:

RAVALLION, M. «The Mystery of the Vanishing Benefits. Ms Speedy Analyst's Introduction to Evaluation». *World Bank Economic Review* (2001), n.º 15, pp. 115-140.

VERA-HERNÁNDEZ, M. «Evaluar intervenciones sanitarias sin experimentos». *Gaceta Sanitaria* (2003), n.º 17, pp. 238-248. (http://scielo.isciii.es/pdf/gs/v17n3/revision.pdf)

ARTÍCULOS QUE REVISAN DIVERSAS TÉCNICAS DE EVALUACIÓN:

BLUNDELL, R.; COSTA DIAS, M. «Evaluation methods for non-experimental data». *Fiscal Studies* (2000), n.º 21 (4), pp. 427-468.

RAVALLION, M. *Evaluating Anti-Poverty Programs*. Washington DC: World Bank, 2006. (Policy Research Working Paper 3625

ARTÍCULOS SOBRE EXPERIMENTOS SOCIALES:

BURTLESS, G. «The case for randomized field trials in economic and policy research». *Journal of Economic Perspectives* (1995), n.º 9, pp. 63-84.

DUFLO, E.; GLENNERSTER, R.; KREMER, M. *Using Randomization in Development Economics Research: A Toolkit*. Londres: CEPR, 2007. (CEPR working paper, number 6059)

ARTÍCULOS SOBRE *MATCHING*:

CALIENDO, M.; KOPEINIG, S. «Some Practical Guidance for the Implementation of Propensity Score Matching». *Journal of Economic Surveys* (2008), n.º 22, pp. 31-72.

IMBENS, G. «Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review». *The Review of Economic and Statistics* (2004), n.º 86, pp. 4-29.

ARTÍCULOS SOBRE VARIABLES INSTRUMENTALES:

HECKMAN, H. «Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations». *Journal of Human Resources* (1997), n.º 32, pp. 441-462.

ARTÍCULOS SOBRE REGRESIÓN DISCONTINUA:

LEE, D.; LEMIEUX, T. *Regression Discontinuity Designs in Economics*. Boston: NBER, 2009. (Working Paper Series, n.º 14723)

ENLACES DE INTERÉS

Network of Networks on Impact Evaluation (NONIE)
http://www.worldbank.org/ieg/nonie/ index.html

Banco de Desarrollo Iberoamericano
http://www.iadb.org/ove/DefaultNoCache.aspx?Action=WUCPublicationsraImpactEvaluations

Evaluaciones de impacto en Colombia
http://www.dnp.gov.co/PortalWeb/Programas/Sinergia/EvaluacionesEstrat%C3%A9gicas/tabid/215/Default.aspx

Evaluaciones de impacto en Chile
http://www.dipres.cl/572/propertyvalue-15223.html

Base de datos del Banco Mundial sobre evaluaciones de impacto
http://web.worLdbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21534261~menuPK:412159~pagePK:210058~piPK:210062~theSitePK:384329,00.html

ivàlua
Institut Català d'Avaluació de Polítiques Públiques
Gran Via de les Corts Catalanes, 680, 3º
08010 Barcelona
T. +34 93 554 53 00
info@ivalua.cat
www.ivalua.cat

Ejemplar impreso: 12 €
Versión electrónica: www.ivalua.cat