Elección del método de evaluación cuantitativa de una política pública

Buenas prácticas en América Latina y la Unión Europea

Ignacio Moral-Arce

Colección **Documentos de Trabajo nº 6**

Serie **Guías y Manuales**
Área **Finanzas Públicas**

EUROsociAL
PROGRAMA PARA LA COHESIÓN SOCIAL EN AMÉRICA LATINA

Ignacio Moral-Arce, en la actualidad coordinador de Área en la Dirección de Estudios del Instituto de Estudios Fiscales, pertenece al Cuerpo Superior de Estadísticos del Estado y ha sido profesor en diversas universidades. Tiene una amplia trayectoria investigadora en mercado laboral y pensiones, y es autor de diferentes trabajos y artículos en revistas internacionales y nacionales sobre salarios, técnicas cuantitativas y mercado de trabajo. Está especializado en evaluación de políticas públicas, tanto desde un punto de vista exante mediante el uso de modelos de simulación, así como evaluación expost. En este último campo ha diseñado y realizado diferentes estudios de evaluación de impacto en políticas de mercado laboral, medio ambiente, así como I+D. Durante los últimos años ha realizado actividades de consultoría y apoyo en temas de evaluación (planificación, diseño e implementación) a los ministerios de Hacienda y Administraciones Públicas, Economía y Competitividad, y Empleo y Seguridad Social en España; y en diferentes países de Latinoamérica como Uruguay, Paraguay, Perú y Ecuador.

Elección del método de evaluación cuantitativa de una política pública

Buenas prácticas en América Latina y la Unión Europea

Ignacio Moral-Arce

Documento de Trabajo nº 6
Serie: Guías y Manuales
Área: Finanzas Públicas

EUROsociAL
PROGRAMA PARA LA COHESIÓN SOCIAL EN AMÉRICA LATINA

Edita:
Programa EUROsociAL
C/ Beatriz de Bobadilla, 18
28040 Madrid (España)
Tel.: +34 91 591 46 00
www.eurosocial-ii.eu

Con la colaboración:
Fundación Internacional y para Iberoamérica
de Administración y Políticas Públicas (FIIAPP)

Instituto de Estudios Fiscales

La presente publicación ha sido elaborada con la asistencia de la Unión Europea. El contenido de la misma es responsabilidad exclusiva de los autores y en ningún caso se debe considerar que refleja la opinión de la Unión Europea.

Edición no venal.

Realización gráfica:
Cyan, Proyectos Editoriales, S.A.

Madrid, enero 2014

Prólogo

"I know thy works, and thy labour, and thy patience."
The Apocalypse Of Saint John (Revelation), capítulo 2

El protagonismo del sector público tiene su base en la necesidad de los ciudadanos de dar respuesta a determinadas necesidades que el sector privado no es capaz de satisfacer por sí solo o lo hace de una manera que podemos considerar injusta o inequitativa. Como bien sabemos, desde el siglo pasado, su actividad se ha ido incrementando paulatinamente, proveyendo a la población de un entramado de servicios que se incorporan en su día a día y que suponen los cimientos de la gobernabilidad y la estabilidad de las democracias más desarrolladas. Ante la convicción de que es necesario contar con un Estado de bienestar sostenible y viable que proporcione a la población bienes públicos y preferentes de calidad, garantes de la igualdad de oportunidades y del bienestar social, se hace imprescindible preguntarnos sobre el papel de las Administraciones públicas en el desempeño de las funciones propias de la Hacienda Pública (asignación, distribución y estabilidad económica). La respuesta pasa indiscutiblemente por la construcción de un sistema eficaz y eficiente en el que los responsables públicos gestionen de manera responsable y transparente los recursos con los que cuentan.

En este sentido, haríamos bien en prestar atención a la reforma que en los últimos años estamos desarrollando en las Administraciones públicas. El cambio de modelo nos acerca hacia una actitud que promueve la importancia de la evaluación de las políticas públicas como instrumento para la consecución de una planificación más racional, transparente y de mayor eficiencia en términos económicos, convirtiéndola en herramienta clave para la toma de decisiones. A través de este manual, el autor, Ignacio del Moral, analiza de manera brillante el diseño y metodología de la función de evaluación. Se trata de una guía metodológica que se ha desarrollado con la pretensión de determinar el diseño de la evaluación cuantitativa óptima, la cual implica la utilización de métodos objetivos, carentes del sesgo evaluador de quien la desarrolla y que se llevan a cabo a través de métodos estadísticos.

Consideramos que es acertado apostar por esta nueva tendencia en la que la evaluación y el seguimiento de las políticas desarrolladas por las Administraciones públicas se muestran como una herramienta útil. Esta nueva disposición está tomando una posición de innegable importancia para la provisión de servicios a la sociedad, por lo que la presente guía es un instrumento de trabajo valioso para controlar la validez y la adecuación de las políticas públicas implantadas con su diseño inicial y con la satisfacción de las necesidades para las cuales fueron elaboradas. Como director del Instituto de Estudios Fiscales me congratula poder aportar trabajos como el presente en el que nos sumamos a las iniciativas de mejora en la gestión pública como fundamento para dar respuesta a los desafíos sociales del siglo XXI.

Prof. Dr. José Antonio Martínez
Director General del Instituto de Estudios Fiscales

Prólogo

La evaluación es una herramienta de enorme potencial para el perfeccionamiento de la política y la gestión públicas. Permite aprender de las políticas, planes o programas analizados; mejorar los futuros; y rendir cuentas a la ciudadanía. Por lo tanto, conduce al fortalecimiento democrático a través de la mejora de la acción y la transparencia del sector público.

La evaluación de impacto es la que representa más claramente la vanguardia de estas técnicas. Por ello, tiene un atractivo especial y se constata un interés creciente en las Administraciones públicas por utilizarla. A pesar de esto, existen muchos retos relacionados con el diseño y puesta en marcha de las evaluaciones de impacto que tienen que ver con la calidad y cantidad de la información disponible o la posibilidad de elección de adecuados grupos de control que condicionan la elección de los métodos más apropiados. El presente manual arroja luces sobre este proceso de elección entre las distintas técnicas con el objeto de conseguir llevar a cabo evaluaciones lo más rigurosas posibles.

Sin embargo, el desafío más importante en relación a las evaluaciones sigue siendo la utilización adecuada de los resultados de las mismas, para lo que se necesita una buena dosis de voluntad política y de capacidad técnica. Los resultados de las evaluaciones de impacto nos permiten identificar y separar las políticas y programas que funcionan y tienen un claro impacto positivo de aquellas que no tienen incidencia en el bienestar de los ciudadanos, con vistas a que puedan ser corregidas o desestimadas.

En América Latina se constata una creciente exigencia de información, tanto por parte de la ciudadanía como de las Administraciones públicas, sobre los resultados y la calidad de los servicios públicos con el fin de conseguir que satisfagan las demandas económicas y sociales. La ampliación de las clases medias en la región ha traído consigo un aumento de la capacidad de supervisión de la sociedad respecto a las actuaciones públicas y una mayor presión para el aumento del refinamiento de las actuaciones del sector público. La instalación definitiva de la evaluación en el ciclo de las políticas públicas puede contribuir a satisfacer esa necesidad, al coadyuvar a una mejor asignación de los recursos públicos aumentando su eficiencia y eficacia.

Por todo esto, EUROsociAL-II apoya la mejora del bagaje técnico de los funcionarios latinoamericanos responsables de llevar a cabo evaluaciones y les acompaña en la realización de algunas evaluaciones piloto que puedan servir como germen para, posteriormente, generalizar el proceso.

De esta forma, el programa busca apoyar la consolidación de una cultura evaluadora en el seno de las Administraciones latinoamericanas; muchas de ellas en procesos ambiciosos de modernización. Esto permitirá a los responsables políticos apostar por la toma de decisiones informadas que mejoren la calidad de los servicios públicos, siendo de especial relevancia para el programa aquellos que impactan más directamente en la cohesión social.

Para ello, la actuación de la Fundación Internacional y para Iberoamérica de Administración y Políticas Públicas (FIIAPP) como socio coordinador del área de finanzas públicas se apoya en la valiosa capacidad técnica del Instituto de Estudios Fiscales de España. Este socio operativo

de EUROsociAL-II es el responsable de la elaboración del presente manual, que se constituye en una potente herramienta para las actividades del programa. Esta línea de impulso de la evaluación va a continuar desarrollándose en EUROsociAL-II, permitiendo extraer lecciones aprendidas de las experiencias de las Administraciones europeas que puedan ser de interés para sus homólogas latinoamericanas, con su lógica adaptación, en el marco de sus procesos de reformas de políticas.

Pedro Flores Urbano
Director de la FIIAPP

1. Introducción

La palabra "evaluación" puede ser interpretada de manera bastante amplia. Significa cosas diferentes para distintas personas y organizaciones. Los ingenieros o los encargados de calidad en un proceso de producción, por ejemplo, pueden evaluar o probar la calidad del diseño de un producto, la durabilidad del material, la eficiencia de un proceso productivo o la seguridad de un puente. Los críticos evalúan la calidad de un restaurante, película o libro. Un psicólogo de niños puede evaluar o valorar el proceso de decisión de los niños. Sin embargo, el tipo de evaluación en que se enmarca este trabajo es el de la evaluación de una política o intervención pública. En este sentido y en pocas palabras, la evaluación de un programa está destinada a responder la pregunta: "¿Cómo está funcionando nuestro programa o política?". Esto puede tener distintas respuestas dependiendo de quién esté preguntando, y a quién le están hablando. Por ejemplo, si un organismo internacional (Banco Mundial, BIRD, etc.) que invierte 10 millones de euros en un programa pregunta al director de dicho programa: "¿Cómo está funcionando nuestro programa?". Esta pregunta se puede interpretar como: "¿Has estado malgastando nuestro dinero?", que, sin duda, puede parecer una especie de interrogatorio. Alternativamente, si un político pregunta a su electorado, "¿Cómo está funcionando nuestro programa?", podría estar simplemente preguntando: "¿Está nuestro programa alcanzando sus metas? ¿Cómo podemos mejorarlo para usted?". Por ende, la evaluación de programas puede ser asociada con sentimientos positivos o negativos, dependiendo de si su objetivo es el de exigir una rendición de cuentas o si se trata de un deseo de aprender.

A un nivel muy básico, las evaluaciones de una política pública (ver Wholey, 2010, López-Acevedo y Tan, 2010 y NSF, 2002 para más detalles) tratan de contestar a la pregunta clave: ¿fue efectivo el programa?, lo que se considera una "evaluación de impacto". Además, en el caso de que el programa estuviera bien pensado en su diseño e implementación, también es posible responder a las siguientes preguntas: ¿cuánto de efectivo fue?, ¿hubo efectos involuntarios?, ¿quién se benefició más?, ¿quién salió perjudicado?, ¿por qué funcionó o por qué no?, ¿qué aprendizajes pueden ser aplicados en otros contextos, o si el programa se lleva a mayor escala?, ¿relación costo-efectividad del programa?, ¿cómo se compara con otros programas diseñados para cumplir los mismos objetivos?

Este documento pretende ofrecer una guía metodológica que permita determinar el diseño de evaluación cuantitativa óptima, en términos de idoneidad, viabilidad y que resulte asequible en términos presupuestarios, dependiendo de las características del programa implementado por el gobierno[1]. Por lo tanto, los dos elementos fundamentales desarrollados en este manual son:

- En primer lugar, identificar criterios para definir si corresponde (o no) hacer una evaluación cuantitativa (y en especial de impacto) en una determinada intervención, estudiando qué condiciones se deben cumplir para que esa política sea merecedora (o no) de una evaluación de este tipo.
- En segundo lugar, después de que se decida realizar una evaluación cuantitativa, definir etapas y criterios que guíen el diseño de la evaluación de impacto más adecuado para la intervención a evaluar, para con posterioridad determinar la técnica de análisis específica.

[1] Es fundamentalmente un documento técnico para personas que se van a encargar de realizar dichas evaluaciones. Para más detalles consultar Gertler et al. (2010), Khandker et al. (2010) y Holden y Zimmerman (2009).

2. Tipos de evaluación de una política pública

Existe la creencia de que la evaluación de impacto lo es todo en la evaluación de una política pública. Sin embargo, la teoría de resultados (o evaluación cuantitativa) coloca a la evaluación de impacto en el lugar que le corresponde como una técnica muy eficaz, pero solamente como una de las posibles formas de ofrecer evidencias empíricas complementarias sobre si la intervención pública funciona como se muestra en Bamberger (1986), Duignan (2009), y Kusek y Rist (2004). Este trabajo no minimiza la importancia y utilidad de la evaluación de impacto, ya que no duda de que pueda proporcionar información única y poderosa sobre la eficacia del programa. Lo que se pretende mediante este documento es poder ofrecer alternativas en aquellas situaciones en las que la evaluación del impacto no es la opción más apropiada, factible o asequible. Para ello, es necesario ayudar a los responsables en la planificación de la evaluación a identificar distintas combinaciones de otros tipos de métodos y pruebas que pueden entrar en juego cuando se realiza la pregunta clave en este campo: "¿Funciona la intervención pública?

Para contestar a esta pregunta existe la posibilidad de emplear evaluaciones de tipo cuantitativo o evaluaciones cualitativas.

- La evaluación cuantitativa está orientada hacia los objetivos que se desean estudiar y aboga por la utilización de métodos cuantitativos, mediante el empleo de métodos estadísticos, para lo que será necesario usar datos "exactos". Los instrumentos empleados serán independientes de sesgos del evaluador.
- La evaluación cualitativa se encuentra más libre de objetivos. Es un enfoque no estructurado, con claro componente "subjetivo" en el que todo conocimiento o información es aceptable. Se encuentra sobre todo orientada a procesos. Desarrolla informes descriptivos, interpretativos o estudio de casos, mediante información "real", "rica" y "profunda".

Además de esta diferenciación entre tipos de evaluación presentada previamente, existen diferentes posibilidades de describir los distintos tipos de evaluaciones que se pueden realizar para estudiar una determinada política o intervención pública. Los más habituales son los siguientes:

2.1. Evaluación de necesidades

Los programas y políticas se realizan para tratar de dar respuesta a unas necesidades específicas que tiene la sociedad. Por ejemplo, podríamos observar que la incidencia de la diarrea en una comunidad es particularmente alta. Esto puede deberse a comida o agua contaminada, mala higiene o cualquier otra explicación plausible. Una evaluación de necesidades puede ayudarnos a identificar la fuente del problema y a aquellos más perjudicados.

La evaluación de necesidades es un enfoque sistemático para identificar la naturaleza y el alcance de un problema social, definir la población objetivo a ser atendida, y determinar la atención que necesitan para hacer frente al problema. Sin duda, una evaluación de necesidades es esencial, porque los programas no serán efectivos si el servicio no se diseña adecuadamente para atender las necesidades o si esas necesidades en realidad ya no existen. Por ejemplo, si las fuentes que contaminan el agua potable están relacionadas con la agricultura, las inversiones en infraestructura de saneamiento, tales como baños y sistemas de alcantarillado, podrían

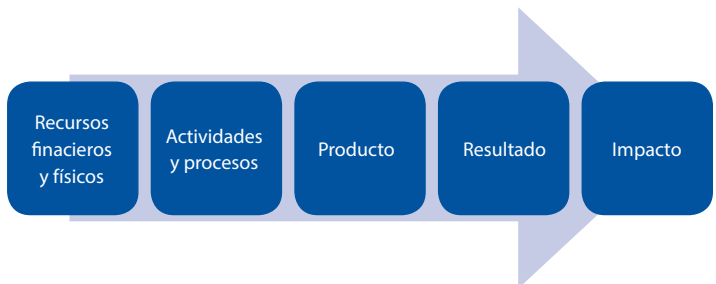
no resolver el problema. La evaluación de necesidades puede ser llevada a cabo utilizando indicadores sociales, encuestas y censos, entrevistas, etc.

2.2. Evaluación teórica y de diseño del programa

Los programas y políticas se realizan para contestar la existencia de ciertas necesidades. Sin embargo, encontrar y solucionar esa necesidad, usualmente, necesita de cierto grado de reflexión. Para los responsables de políticas públicas requiere la identificación de las razones que causan esos resultados indeseables, y elegir aquellas estrategias (entre una larga lista de opciones) para lograr tener distintos resultados.

Una evaluación teórica del programa trata de modelizar la teoría que está detrás del programa, presentando un plan viable y factible para mejorar la situación de los individuos. Si las metas y supuestos en los que se basa son irracionales, entonces existen muy pocas posibilidades de que el programa sea efectivo. La evaluación teórica del programa incluye, primero, articular el programa teórico y, después, evaluar cómo de bien la teoría responde a las necesidades de la población objetivo. Las metodologías usadas en la evaluación teórica de programas incluyen el Enfoque del Marco Lógico o Teoría del Cambio. En la siguiente figura se muestra un ejemplo simple de un marco lógico:

Figura 1. Marco lógico de una política pública



2.3. Evaluación de procesos

Antes de ser lanzado, cualquier programa existe a nivel conceptual, pero una vez implementado, el programa se enfrenta a la realidad del terreno, y comienzan las preguntas del tipo: ¿la organización cuenta con un equipo bien entrenado?, ¿están asignadas las responsabilidades de forma correcta?, ¿están siendo completadas las tareas de los organismos intermediarios a tiempo?

La evaluación de procesos, también conocida como evaluación de la implementación, analiza la efectividad de las operaciones del programa, la implementación y la entrega de producto y de servicios. La evaluación de procesos nos ayuda a determinar, por ejemplo:

- Si los servicios y metas están alineados apropiadamente.
- Si los productos o servicios están siendo entregados a los destinatarios, como se pretendía.
- Cómo de bien está organizado el servicio de entrega.

- La efectividad de la gestión del programa.
- El grado en que se están usando los recursos del programa.

Las evaluaciones de procesos son usadas a menudo por los administradores como puntos de referencia para medir el éxito.

2.4. Evaluación de impacto

El principal propósito de una evaluación de impacto es la determinar si un programa tiene impacto (en unos cuantos resultados clave), y más específicamente, cuantificar *qué grande* es este impacto. La primera pregunta que nos tenemos que realizar es: ¿qué es impacto?, y es aquella variable de interés de largo plazo que queremos modificar: reducción de la criminalidad, aumento del nivel de renta de la población, reducción de tasas de mortalidad, incremento de tasas de escolaridad en la población infantil, etc. En el ejemplo de reducción de tasas de mortalidad, el impacto de la política consiste en ver cuánto más saludable están las personas gracias al programa en comparación a la situación en la que podrían haber estado sin él. O más específicamente, cuánto disminuyó la incidencia de la enfermedad con el programa en comparación a lo que hubiera ocurrido si no hubiera existido.

Conseguir esta medición de manera correcta es más difícil de lo que parece. Es posible medir la incidencia de una determinada enfermedad en una población que recibe el programa, pero es imposible medir directamente cómo hubiera estado esta misma población si no hubiese recibido el programa —así como es imposible medir cuál sería la enfermedad más mortal hoy en día si no se hubiese descubierto la penicilina, ya que es posible que pequeñas heridas siguieran siendo causantes de muchas muertes, o alternativamente, algo parecido a la penicilina hubiese sido descubierto en un laboratorio diferente en otra parte del mundo—.

Las evaluaciones de impacto, usualmente, estiman la efectividad de un programa al comparar los resultados de aquellos (individuos, comunidades, escuelas, etc.) que participaron en el programa frente a los que no lo hicieron. El desafío clave en una evaluación de impacto es el encontrar un grupo de personas *que no participaron*, pero que son lo suficientemente parecidas a las que se beneficiaron del programa como para medir "cómo estarían los participantes si no hubiesen recibido el programa". Hay varios diseños para hacer esto y cada diseño viene acompañado de sus propios supuestos[2]. Para analizar en profundidad este enfoque de evaluación los trabajos de Ravillion (2008) y Heckman y Vytlacil (2005) detallan de manera analítica una gran cantidad de las posibilidades de aplicación.

2.5. Evaluación costo medio, costo-beneficio y costo-efectividad

Es probable que diferentes organizaciones o departamentos tengan estrategias muy distintas a la hora de enfrentarse a un mismo problema. Supongamos el caso de que el suministro de agua de una comunidad estuviera contaminado generando una epidemia de diarrea en la población.

[2] Dentro de la evaluación de impacto, la primera gran diferencia que se establece es entre evaluación ex-post y evaluación ex-ante. La primera de ellas es la evaluación que se realiza cuando el programa público ya se ha realizado o se está implementando utilizando datos observados. Sin embargo, la evaluación de impacto ex-ante se centra en estimar el efecto de la política antes de que esta se realice, realizando previsiones o proyecciones utilizando modelos de microsimulación.

Es posible que una determinada ONG pueda abogar por realizar inversiones en infraestructuras para lograr sanear el agua, mediante un sistema de alcantarillado, tuberías de agua, etc. Otra ONG podría proponer un sistema de distribución donde los hogares reciban, gratuitamente, tabletas de cloro para tratar el agua en su propia casa. Si estos dos métodos fuesen igualmente efectivos —ya que cada uno de ellos es capaz de reducir la diarrea en un 80%—, la pregunta siguiente que surge es ¿los responsables políticos estarían igual de contentos implementando una u otra política? Probablemente no; ya que aunque tienen el mismo grado de efectividad sería necesario considerar los costes de cada estrategia.

Es muy probable que la inversión en infraestructura en un pueblo lejano sea excesivamente cara. En este caso, la opción sería clara. No obstante, las opciones no son siempre tan obvias. Una opción más realista (pero aún hipotética) sería comparar una inversión en infraestructuras que reduce la diarrea en un 80%, frente a un programa de distribución de tabletas de cloro que cuesta 50 veces menos, pero que solo reduce la diarrea en un 50%. En esta situación, ¿con qué opción se quedaría el responsable político?

Un análisis costo-beneficio cuantifica los beneficios y costes de una actividad y los pone en la misma medida métrica (a menudo en una unidad monetaria). Se trata de responder la pregunta: ¿está el programa produciendo suficientes beneficios para compensar los costes? O en otras palabras, ¿la sociedad será más rica o más pobre después de realizar esta inversión? De todas formas, tratar de cuantificar el beneficio de la salud de los niños en términos monetarios puede ser extremadamente difícil y subjetivo. Por lo tanto, cuando el valor exacto del beneficio carece de un amplio consenso, este tipo de análisis puede producir resultados que son más controvertidos que esclarecedores. Este enfoque es más útil cuando hay múltiples tipos de beneficios y se ha acordado monetizarlos.

3. Beneficios de una evaluación cuantitativa de una política pública

El análisis de la evaluación cuantitativa es un importante tipo de análisis que a menudo se promueve como una forma de ayudar a las personas encargadas de tomar decisiones para poder seleccionar entre diferentes tipos de intervención. Este enfoque permite realizar dos contribuciones importantes a la teoría de la evaluación cuantitativa de una intervención pública. La primera es la de insistir en una mayor transparencia en torno a la inclusión de costes y beneficios en todos los análisis económicos. El segundo es proporcionar reglas de decisión acerca de los posibles tipos de evaluación cuantitativa que se podrían utilizar en un caso particular, sobre la base de qué tipo de estimaciones se obtienen en la evaluación de impacto en cada caso concreto. Además, realizar una evaluación cuantitativa suele requerir conocimientos especializados técnicos, así como disponer de ciertos expertos que es necesario consultar. La intención de este documento no es explicar cómo se debe realizar una evaluación, sino proporcionar una guía sencilla para determinar qué tipos de evaluación son adecuados en cada caso.

3.1. Necesidad de mayor transparencia de los impactos, costes y beneficios en las evaluaciones cuantitativas

Una persona sin conocimientos de economía puede tener dificultades si intenta leer un análisis de impacto, o de costo-beneficio o costo-eficacia que le permita determinar con rapidez qué costes y qué beneficios se han incluido en el análisis. Es posible que existan costes o beneficios realmente importantes, pero que han sido excluidos del análisis cuantitativo. En estas situaciones resulta imprescindible ser prudente a la hora de utilizar estos resultados en la toma de decisiones futuras.

Una forma eficaz de comunicar de manera rápida y exacta el tipo de costes y resultados que se han introducido en el análisis cuantitativo, es mostrar las variables de la intervención mediante un modelo visual, como son los marcos lógicos. Estos modelos muestran en un formato esquemático todos los resultados a largo plazo (es decir, nuestra variable de interés que es el impacto del programa) que se buscan por una intervención pública, así como todos los pasos (variables) de nivel inferior, ya sean estos resultados a medio y corto plazo, que se consideran necesarios para alcanzar el objetivo final a largo plazo. Estos modelos se diseñan de acuerdo con un conjunto de reglas, que garantizan que los resultados de una determinada intervención pública representan con exactitud sus medidas y resultados. Es importante que el modelo de resultados de una intervención represente un panorama completo de lo que se cree que va a ocurrir en la intervención. Por ejemplo, el marco lógico debe indicar si los resultados de la variable de interés son cuantificables o no, si están controlados completamente por la intervención, o si los efectos son buscados o no (externalidades del programa), entre otro tipo de información.

Sin duda, para evaluar de forma óptima, es necesario incluir un marco lógico en el inicio de cada análisis de impacto, costo-eficacia o costo-beneficio para que se puedan determinar las variables que se utilizan en el análisis. Si esto se hace de manera sistemática en todas las evaluaciones cuantitativas entonces es posible ofrecer a cualquier lector una herramienta útil que permita

determinar rápidamente qué variables están (y cuáles no) incluidas en el análisis[3]. Además, el uso habitual de estos "modelos de resultados" en los análisis cuantitativos también facilita la labor a las personas que desean comparar diferentes informes de impactos, costes y beneficios de una misma intervención pública, ya que se puede comprender rápidamente si las diferencias en los resultados de los diferentes análisis se deben a incluir o no determinadas variables en el estudio.

3.2. Transparencia respecto a si la medición del efecto de la política tiene fundamentos empíricos o no

La segunda cuestión en la que los "modelos de resultados" abogan por una mayor transparencia en el análisis cuantitativo hace referencia a la credibilidad (o valor probatorio) de las estimaciones realizadas sobre los efectos de la política. Desde el punto de vista de la teoría de los resultados, el efecto de una intervención se define formalmente como la cantidad de cambio en la variable de interés (resultado a largo plazo o impacto) que se puede atribuir completamente al efecto causal de etapas intermedias (resultados a medio o corto plazo) o un conjunto de pasos —es decir, una intervención— dentro del mismo modelo de resultados.

El análisis tradicional de costo-efectividad y costo-beneficio tiende a prestar poca atención a la cuestión de con qué calidad se calculan las estimaciones del efecto en los análisis económicos, porque el énfasis se centra en la metodología empleada en el análisis de la rentabilidad del mismo (estudio de costes y beneficios), examinando características tales como qué tasa de descuento es necesaria aplicar. Sin embargo, la cuestión que a menudo no se contempla de manera integrada, y que es fundamental, es la relativa a la base empírica que se ha empleado para calcular el efecto de una intervención, que es un elemento clave para determinar las estimaciones de beneficios futuros.

Por supuesto, la eficacia, la rentabilidad y los análisis de costo-beneficio dependen considerablemente de la exactitud de estos valores, dado que un error significativo de estas estimaciones puede hacer que cualquier análisis de costo-eficacia o de costo-beneficio sea completamente inútil (ya que proporciona una sensación de falsa seguridad para la toma de decisiones debido a que la información utilizada para decidir presenta grandes deficiencias), y la falta de disciplina y transparencia en el logro de tales estimaciones simplemente anima la polémica basada en que es posible cualquier resultado que un individuo (ya sea este un político, gestor de proyecto, organización) desee obtener solamente mediante la combinación y elección de diferentes estimaciones de las variables clave de la intervención pública. Por ello, las críticas que se realizan sobre la "calidad" de las estimaciones en una evaluación cuantitativa (sobre todo en las económicas) se consideran habitualmente como una mera cuestión técnica y algo residual, dado que la mayoría de las personas se centran exclusivamente en utilizar los resultados del análisis y no tanto cómo se llega a ellos.

Un aspecto clave de la teoría de los resultados es el relativo a la información disponible para realizar un estudio cuantitativo, es decir: lo que se sabe y lo que no se sabe sobre los efectos de las intervenciones, que es primordial a la hora de realizar el estudio de la evaluación de impacto, tanto cuando disponemos de toda la información, así como en muchos casos en los que no se

[3] El intento de comunicar esta información en una forma narrativa es mucho más ineficiente que el uso de un enfoque visual.

tiene información completa. Fundamentalmente, este enfoque intenta transmitir de manera sencilla a los responsables en qué situación se encuentran para la toma de decisiones —un entorno de información de alta calidad o, por el contrario, un ambiente donde hay poca información sólida acerca de los efectos del programa—[4]. Llegado este punto, resulta fundamental tener claro los diferentes tipos de evaluaciones cuantitativas que se pueden realizar:

- Análisis cuando "no" se dispone de información de resultados para realizar una estimación sobre el efecto del impacto, es decir, no se tiene información de nuestra variable de interés a largo plazo (indicador de impacto), y tampoco de indicadores a corto y medio plazo (en el mejor de los casos se tiene información de indicadores de productos).
- Análisis cuando se dispone de información empírica para estimar el impacto mediante el efecto que la política tiene sobre variables de resultados a medio o corto plazo
- Análisis cuando se dispone de información empírica para realizar la estimación del impacto en una variable de resultados a largo plazo (una variable de impacto).

Teniendo presente estas tres opciones, que dependen de la información disponible, la situación correcta en la que se pueden realizar evaluaciones de impacto es la tercera, ya que las dos condiciones clave para realizar un estudio de esta característica es disponer de información de la variable de interés y que esta variable refleje el comportamiento a largo plazo buscado por la política que se ha implementado. En el caso de disponer de información de medio y corto plazo se puede estudiar el efecto inmediato (efecto coyuntural) que tiene una política, pero no se puede decir nada sobre el largo plazo (efecto estructural), que es precisamente el objetivo. Finalmente, si no se puede disponer de información de resultados de ningún tipo, resulta inviable realizar ningún tipo de estudio de impacto de una intervención pública.

[4] En la actualidad, el medio empleado para analizar la incertidumbre en las estimaciones es a través de un análisis de sensibilidad. Un análisis de sensibilidad consiste en realizar un análisis económico una serie de veces variando aquellas estimaciones clave sobre las que se cree que existen dudas con el fin de ver el efecto que esta modificación puede tener en el resultado final.

4. Evaluación cuantitativa: situaciones en que debe (o no) utilizarse

La evaluación cuantitativa de una política pública siempre debe tenerse en cuenta al planificar la evaluación de un determinado programa o intervención pública. Es un tipo de evaluación muy eficaz, ya que, cuando se hace correctamente, permite establecer qué cambios en nuestra variable de interés (variable de resultado a largo plazo) son atribuibles a un programa en particular. Sin embargo, suponer que siempre, o en la gran mayoría de casos, se debe realizar una evaluación de impacto es un error. La evaluación de impacto trata de medir (aislar) qué porcentaje de la variación en nuestra variable objetivo se debe a la realización de una determinada intervención. El primer elemento a considerar es que, a veces, este término se utiliza para contrastar la evaluación de resultados a corto plazo, en lugar de los efectos reales de una intervención sobre los resultados a más largo plazo dentro del modelo de resultados de una intervención. Para solventar este problema, disponer de un modelo visual de los resultados de alto nivel y todos los pasos que conducen a ellos puede resolver considerablemente la situación.

Es necesario tener claro que la atribución de los efectos en la variable de resultados a largo plazo (es decir, el impacto) de una intervención pública es uno de los pilares básicos de todos los sistemas basados en resultados, pero es necesario volver a recordar que la evaluación cuantitativa de una política puede ser cualquier tipo de evaluación, como la evaluación de seguimiento, gestión del rendimiento o planificación estratégica, que trata de especificar los resultados de un programa, medirlos, atribuirlos, y responsabilizar a las resultados a corto y medio plazo para dar cuenta de su consecución. Sin duda, esto es un enfoque mucho más amplio que el dado exclusivamente por las evaluaciones de impacto, y el objetivo de este documento consiste en determinar si realizar una evaluación de impacto es lo más adecuado, si es viable y asequible, cuando se pretende estudiar la eficacia de una determinada intervención pública.

4.1. El problema de suponer que siempre se puede realizar una evaluación de impacto

No se debe suponer, antes de tiempo, que siempre se debe realizar una evaluación de impacto de un programa. Es necesario saber, para cada caso, si se va a realizar una evaluación de impacto, empleando para ello un análisis cuidadoso de la pertinencia, la viabilidad y la factibilidad (económica) de la implementación de esta evaluación, siendo estos tres aspectos de vital importancia, ya que la evaluación de impacto puede, o no, ser apropiada, factible y/o económicamente viable dependiendo de cada tipo de programa o intervención pública que se desea estudiar.

Por desgracia, muchos interesados, a distintos niveles, creen que se puede (y se debe) realizar evaluaciones de impacto de manera rutinaria a todos los programas. Paradójicamente, esa insistencia en tratar de realizar evaluaciones de impacto de manera sistemática, puede conducir al resultado no deseado de desperdiciar recursos (que son limitados) por tratar de realizar una evaluación de este tipo. La insistencia ingenua de que solo los resultados de la evaluación de impacto se deben utilizar para determinar qué intervenciones deben realizarse, sin tener en cuenta la conveniencia, viabilidad y asequibilidad económica de dicha evaluación de impacto puede provocar graves distorsiones de los sistemas de monitoreo, seguimiento y evaluación, ya que pueden llevar a la situación extrema de que solo se termine haciendo aquello que se puede evaluar fácilmente. Por el contrario, las decisiones estratégicas no deben basarse solo en hacer

aquello que es fácilmente evaluable, sino hacer aquello que tenga la mayor posibilidad de éxito y que sea lo más eficaz posible.

4.1.1. Creer que siempre se debe hacer una evaluación de impacto genera problemas

Es posible que se crea que para cualquier programa público se debe realizar una evaluación de impacto. Este enfoque puede provocar la aparición de una serie de problemas. Estos son:

- La búsqueda ciega y a cualquier precio de la evaluación de impacto. Situaciones en las que la evaluación no es, ni apropiada, ni factible, o muy costosa, llevan a realizar pseudoevaluaciones de impacto de baja calidad, que emplean métodos estadísticos y econométricos técnicamente deficientes, y que pueden convencer a ciertos políticos o algunos grupos de interés de que se ha realizado un estudio de impacto correcto, pero que, al ser sometidas a una revisión metodológica completa por parte de personas expertas en la materia no serán aceptadas.
- Pérdida de recursos de evaluación. Los recursos de evaluación se deben asignar de manera estratégica para que no se desperdicien. Es posible que la mejor estrategia para ahorrar los escasos recursos de evaluación consista en llevar a cabo evaluaciones de impacto bien diseñadas y bien ejecutadas exclusivamente en algunos casos, y no realizar este tipo de evaluaciones para muchas políticas publicas. También puede suponer un uso mucho más eficaz de los recursos existentes para realizar evaluaciones, el hecho de efectuar evaluaciones formativas (es decir, que contribuyan a asegurar la aplicación eficaz) o realizar evaluaciones de proceso (aquellas que sirven para ayudar a describir un programa ya existente, de modo que se puedan transmitir estas "buenas prácticas" a otros tipos de programas) y no utilizar esos recursos escasos en la realización de la evaluación de impacto.

4.2. Determinar cuándo se debe realizar una evaluación de impacto

El siguiente proceso sirve para determinar si se debe llevar a cabo una evaluación de impacto.

1. ¿Las prioridades de evaluación de un determinado sector coinciden con las de la evaluación de la intervención pública? Esta aproximación cambia el enfoque de la planificación de la evaluación de impacto, ya que en lugar de estar "centrada en el programa" presenta un enfoque "centrado en el sector". En un enfoque de evaluación "centrado en el programa", la pregunta fundamental que se realiza es cómo se debe evaluar de manera correcta un programa. Sin embargo, un enfoque "centrado en el sector" se inicia con la pregunta: "¿cuáles son las necesidades de conocimientos estratégicos del sector en el que se encuentra este programa?". Una vez que se ha obtenido una comprensión clara de esto, entonces, se puede trabajar de nuevo en determinar cuál es el mejor uso de los recursos de evaluación en relación con el programa específico. Sin duda, considerar inicialmente cuáles son las necesidades de información estratégica de un determinado sector y lo que esto puede significar para la evaluación de un programa en particular es mejor que solo planificar una evaluación desde el punto de vista demasiado simplista "centrado en el programa".
2. Determinar si se debe realizar una evaluación de impacto de un programa piloto o de todo el programa. Es importante distinguir entre estos dos enfoques. La segunda aproximación

se da cuando se lleva a cabo la evaluación de impacto de la "plena implementación nacional" del programa. En este caso, la pregunta de evaluación que se plantea es: "¿la plena puesta en marcha del programa mejora los resultados a largo plazo? Una opción alternativa consiste en hacer solo la evaluación de impacto en un programa piloto y entonces, si este tiene éxito, ya no resultará necesario realizar una evaluación de impacto para el pleno despliegue del programa, dado que ya sabemos los efectos beneficiosos que va a generar el programa. En estos casos, todo lo que se hace cuando se realiza la implementación completa del programa a nivel nacional es supervisar la aplicación de "buenas prácticas" observadas en el programa piloto. Las preguntas de evaluación que se deben plantear y contestar en este primer enfoque son: 1) "¿el programa piloto mejora realmente la variable de resultados de largo plazo?" (una pregunta de evaluación de impacto), 2)"¿cuáles son los detalles más relevantes de la implementación del programa piloto?" (una pregunta de evaluación de proceso) y 3) "¿se han aplicado las 'buenas prácticas' observadas en el programa piloto en la implementación completa a nivel nacional del programa?" (una pregunta de evaluación formativa).

Un buen ejemplo de la utilización generalizada de este primer paradigma de la evaluación de un programa piloto se encuentra en el área de la medicina mediante el estudio de los tratamientos médicos. Cuando un paciente acude a un médico y se le prescribe un tratamiento farmacológico, a menudo no se lleva a cabo ninguna evaluación de impacto para determinar si la mejora en el paciente se produce debido al tratamiento, placebo u otro factor. Sin embargo, la clave de todo el enfoque consiste en que el médico está aplicando en su decisión de "dar el tratamiento" criterios de "buenas prácticas" con base en evaluaciones de impacto previas que han sido llevadas a cabo en una fase "piloto" (es decir, en el curso de los ensayos de medicamentos).

Sin lugar a dudas, los estudios en medicina de los distintos tratamientos médicos se ven como una tarea relativamente centrada en pruebas. Esto permite alentar a los programas sociales a adoptar un enfoque más basado en la evidencia, de tal forma que la medicina clínica sirva como ejemplo a seguir. Sin embargo, a veces se realizan intentos ingenuos de estudiar los programas sociales utilizando evaluaciones de impacto para programas que han sido totalmente implementados, situación en la que, claramente, esto no es ni lo más apropiado, ni factible o asequible. Si quienes diseñan la evaluación de dichos programas sociales tienen claras las diferencias entre los dos paradigmas descritos, entonces se debería utilizar la evaluación del programa piloto, que es una aproximación mucho más apropiada, y posteriormente utilizar los resultados obtenidos de "buenas prácticas" para el seguimiento y monitoreo de la plena implementación del programa. Al hacerlo de este modo, se está emulando la aproximación que se utiliza en el tratamiento médico.

3. Analizar la adecuación, viabilidad y asequibilidad de los distintos diseños de evaluación cuantitativa. Si, teniendo en cuenta los dos anteriores puntos, las autoridades han decidido llevar a cabo una evaluación de impacto de un programa, es necesario considerar tres factores que siempre se deben tener presente, que son su conveniencia, viabilidad y asequibilidad de los diferentes diseños de evaluación.

   - La idoneidad o conveniencia se refiere a cuestiones tanto éticas como culturales que están relacionadas con la realización de una evaluación de impacto.
   - La factibilidad se refiere a si va a ser posible llevar a cabo de manera real y efectiva la evaluación de impacto. Tal vez este aspecto sea el más difícil de determinar, ya que por

lo que se refiere a esta característica, es necesario asegurarse de que la evaluación de impacto va a generar una conclusión, así, a veces, las evaluaciones de impacto son supervisadas por personal inexperto en estas tareas, y existe la posibilidad de que puedan surgir muchos problemas prácticos a lo largo del curso de una evaluación de impacto que supongan el abandonado de esta antes de que se puedan proporcionar resultados útiles.

- Por último, incluso en el caso de que la evaluación de impacto sea apropiada y factible, puede no ser asequible. La asequibilidad debe tenerse en cuenta en relación con los posibles usos alternativos de los recursos de evaluación que se utilizarán en la evaluación de impacto, dado que existe la posibilidad de que la realización de una evaluación de impacto realmente efectiva pueda ser un ejercicio muy costoso.

4.2.1. La decisión de emplear un diseño de evaluación de impacto

Ya se ha comentado que suponer que siempre se debe realizar una evaluación de impacto es un error, ya que pueden existir otros diseños de evaluación más adecuados. Partiendo de la información dada en la sección 2 de este documento, las formas de evaluación de una política pública se pueden resumir en tres. La primera es la evaluación formativa que pretende ayudar a optimizar la implementación de la intervención pública, este tipo de evaluación utiliza una serie de técnicas de evaluación tales como modelos lógicos o consultas con los interesados y análisis de necesidades para asegurarse de que el programa o intervención tiene las mayores posibilidades de éxito. La segunda es la evaluación de proceso, que tiene como objetivo describir el curso y el contexto de un programa o intervención, este tipo de evaluación ayuda en la interpretación de los resultados de la evaluación de impacto, y se puede introducir en la evaluación formativa para mejorar el programa. También vale para identificar "buenas prácticas", que se pueden utilizar para mejorar otros programas en el futuro. El tercer tipo es la evaluación de impacto que trata de atribuir los cambios en la variable de resultado (previsto, no, positivo y negativo) a un determinado programa o intervención. Teniendo todo esto en cuenta, a menudo es mucho más estratégico establecer una evaluación que utilice la evaluación formativa, que permita asegurar que la intervención se lleva a cabo de una manera óptima, y utilizar una evaluación de procesos para identificar las mejores prácticas, y no realizar una evaluación de impacto, debido a que esta última evaluación a menudo es costosa, y asegurarse de que se lleva a cabo de manera correcta puede resultar muy difícil.

4.2.2. El diseño de evaluación de impacto puede ser una decisión técnicamente compleja

En aquellas situaciones en las que se ha decido realizar una evaluación de impacto, el siguiente paso consiste en decidir qué diseño entre todos los posibles se debe emplear, lo que puede llegar a ser un ejercicio muy técnico. Sin lugar a dudas, para ello, es imprescindible que el planificador de la evaluación deba estar familiarizado con todos los posibles diseños de evaluación de impacto existentes. El propósito de lo que queda de documento es proporcionar un marco de decisiones a los diseñadores de una evaluación, para que puedan trabajar con una base metodológica sobre diseños de evaluación de impacto y poder justificar por qué se ha tomado una determinada decisión sobre qué técnica de estimación utilizar. En el caso de que la persona que planifica la evaluación no esté familiarizada con los diseños de evaluación de impacto es recomendable que se ponga en contacto con personas que tengan más experiencia en este tema. Sin embargo, si trabajan de forma continuada con información como la dada en este documento, debería permitirles, a través del uso repetido de estas tablas, poder tomar decisiones, dado que poseen un marco más coherente y sólido sobre posibles diseños de la evaluación de impacto.

5. Tipos de técnicas de evaluación de impacto

La pregunta básica de evaluación de impacto constituye esencialmente un problema de inferencia causal. La evaluación de impacto de un programa equivale a evaluar el efecto causal del programa en función de los resultados obtenidos. La mayoría de las cuestiones de política implican relaciones de causa y efecto: ¿mejora la formación del profesorado las calificaciones de los estudiantes?, ¿los programas de transferencias monetarias condicionadas producen mejores resultados de salud en los niños?, ¿los programas de formación profesional aumentan los ingresos futuros de los estudiantes?

Aunque las preguntas de causa y efecto son bastante comunes, no es fácil determinar que una relación entre dos variables sea de causalidad. En el contexto de un programa de formación profesional, por ejemplo, la simple observación de que el ingreso de un individuo aumenta después de que él o ella hubieran completado un programa de capacitación no es suficiente para establecer la causalidad. El ingreso de una persona podría haber aumentado incluso si no hubiera realizado el curso de capacitación debido a su propio esfuerzo, a las condiciones cambiantes del mercado de trabajo, o por otro tipo de factor que puede afectar a los ingresos. Las evaluaciones de impacto nos ayudan a superar el reto de estudiar la causalidad empíricamente al establecer en qué medida un determinado programa, y solamente ese programa, contribuyó al cambio en la variable de resultado. Para establecer la causalidad entre un programa y un resultado, utilizamos los métodos de evaluación de impacto que nos permiten descartar la posibilidad de que la variable de interés se viera afectada por otros factores, aparte del programa de interés.

La evaluación de impacto trata de dar respuesta a la siguiente pregunta, ¿cuál es el impacto o efecto causal de un programa de "P" en un resultado de interés "Y" que se puede expresar mediante la fórmula de evaluación de impacto?:

$$\alpha = (Y \mid P = 1) - (Y \mid P = 0) \qquad (1)$$

Esta fórmula nos indica que el impacto causal de un programa (P) en un resultado (Y), denominado por α, es la diferencia entre la variable de resultado (Y) cuando el individuo recibe el programa (en otras palabras, cuando P=1) menos el resultado Y en el caso de no recibir el programa (es decir, cuando P=0). Por ejemplo, supongamos que P es un programa de capacitación y queremos ver su efecto sobre la variable de resultados Y que es el ingreso de esa persona. El impacto causal del curso de capacitación (α) es la diferencia entre el ingreso que la persona (Y) tendría tras realizar el curso (P=1) menos el ingreso (Y) que tendría esa misma persona, y en el mismo momento de tiempo, en el caso de no haber realizado el curso (P=0).

En otras palabras, el investigador quiere medir el ingreso en el mismo momento de tiempo para la misma unidad (en este caso una persona) pero en dos estados distintos. Si esto fuera posible, observaríamos qué ingreso tendría un mismo individuo si hubiera hecho el curso y si no lo hubiera cursado, de tal modo que la única explicación sobre la diferencia de ingresos es debida a la realización del curso. Mediante la comparación de la misma persona consigo misma en el mismo momento, habríamos logrado eliminar cualquier factor externo con el que también podríamos haber explicado la diferencia en los resultados. Entonces, es posible estar seguros de que la relación entre el programa de formación profesional y los ingresos es causal.

La fórmula de evaluación de impacto presentada en (1) es válida para cualquier política o intervención pública que se desee analizar mediante el estudio de una persona, una familia, una comunidad, una empresa, una escuela, un hospital, o cualquier otra unidad de observación que puede recibir o ser afectado por un programa, así como para cualquier variable de resultado (Y) que está plausiblemente relacionada con el programa en cuestión. Una vez que medimos los dos componentes clave de esta fórmula, el resultado (Y), con y sin el programa, entonces es posible responder a cualquier pregunta sobre el impacto del programa.

5.1. El contrafactual

Como se expresó anteriormente, se puede pensar que el impacto de un programa (α) se obtiene como la diferencia de la variable de resultados (Y) para el mismo individuo en los casos de recibir y no recibir un programa. Sin embargo, la medición en la misma persona de dos estados diferentes al mismo tiempo es imposible, ya que un individuo participó o no en el programa, pero no se dan ambos resultados a la vez, es decir: la persona no puede ser observada simultáneamente en dos estados diferentes (en otras palabras, con y sin el programa). Esto se conoce como "el problema contrafactual": ¿cómo se mide lo que hubiera pasado si la otra circunstancia hubiera prevalecido? Aunque se puede observar la variable de resultado (Y) para los participantes en el programa (Y|P=1), no hay datos que nos digan cuál hubiera sido el valor de su variable de resultado en el caso de no haber recibido el programa (Y|P=0), y es precisamente este término (Y|P=0) el que representa al contrafactual. En otras palabras, se puede pensar que esa cantidad nos está diciendo qué hubiera ocurrido si el participante no hubiera participado. Es decir, indica el valor de la variable de resultado (Y) en el caso de ausencia de un programa (P).

Por ejemplo, supongamos un "niño A " que recibe una vacuna y luego muere cinco días después. El hecho de que el niño A muera después de recibir una vacuna no puede concluir que la vacuna causó la muerte. Tal vez el niño estaba muy enfermo cuando recibió la vacuna, y fue la enfermedad más que la vacuna la que le causó la muerte. Inferir la causalidad entre vacuna y muerte (o enfermedad) va a requerir descartar otros posibles factores que puedan afectar al resultado en cuestión.

En este ejemplo simple de determinar si recibir una vacuna causa la muerte al niño A, un evaluador tendría que establecer qué habría pasado con el niño A en el caso de no haber recibido la vacuna. Dado que el niño A de hecho recibió la vacuna, no es posible observar directamente lo que habría pasado si no la hubiera recibido. "¿Qué le hubiera pasado de no haber recibido la vacuna?" es la situación hipotética, y, a la vez, el principal desafío al que se enfrenta un evaluador, por lo que el elemento clave para realizar la evaluación de impacto consiste en determinar un estado "contrafactual" lo mejor posible para ver qué valor toma esta variable de resultado.

Por lo tanto, al llevar a cabo una evaluación de impacto, es relativamente fácil obtener el primer término de la fórmula dada en (1) que es (Y|P=1)-el resultado en tratamiento (es decir, medir el resultado de interés para la población que participó en el programa). Sin embargo, el segundo término de la fórmula (Y|P=0) no puede ser observado directamente en los participantes del programa, de ahí, la necesidad de llenar este elemento faltante de (1) mediante información que permita obtener una estimación del contrafactual. Para ello, usamos normalmente un grupo de comparación (a veces llamado "grupo de control").

5.2. La estimación del contrafactual

Para ilustrar aún más la estimación del contrafactual, pasamos a un ejemplo hipotético, que ayudará a pensar a través de este concepto clave un poco más a fondo. A nivel conceptual, la solución del problema contrafactual requiere que el evaluador pueda identificar un "clon perfecto" (o réplica perfecta) para cada participante en el programa. Por ejemplo, digamos que el señor B recibe una transferencia del gobierno de 20 euros (política P), y queremos medir el impacto que esta política tiene en su consumo de manzanas (variable Y). Si se pudiera identificar un clon perfecto para el señor B, la evaluación sería fácil: solo con comparar el número de manzanas que comió el señor B (digamos, 6) respecto al número de manzanas que comió su clon (por ejemplo, 4). Por lo tanto, aplicando la ecuación (1) se obtiene:

$$\alpha = (Y \mid P = 1) - (Y \mid P = 0) = 6-4 = 2$$

En este caso, el impacto de la transferencia de dinero sería la diferencia entre esos dos números: 6-4= 2. Sin embargo, en la práctica sabemos que es imposible identificar réplicas perfectas: incluso entre gemelos genéticamente idénticos, hay diferencias importantes.

Aunque no existe un clon perfecto para una sola persona, se pueden usar herramientas estadísticas que permitan generar dos grupos de individuos que, si sus tamaños muestrales son lo suficientemente grandes, son estadísticamente indistinguibles entre sí, dado que no es posible observar diferencias significativas entre ambos. En la práctica, un objetivo clave de una evaluación de impacto es identificar un grupo de participantes en el programa (grupo de tratamiento) y un grupo de no participantes (grupo de control) que son estadísticamente idénticos en ausencia del programa. Si los dos grupos son iguales, con la única excepción de que un grupo participa en el programa y el otro no, entonces podemos estar seguros de que cualquier diferencia en los resultados es debido al programa.

Sin lugar a dudas, el elemento clave, entonces, es identificar un grupo de comparación o control válido que tenga las mismas características que el grupo de tratamiento. En concreto, los grupos de tratamiento y control deben ser similares al menos en tres elementos:

- En primer lugar, el grupo de tratamiento y el grupo de comparación deben ser idénticos en ausencia del programa. Aunque no es necesario que cada unidad en el grupo de tratamiento sea idéntica a cada unidad en el grupo de comparación, en promedio, las características de los grupos de tratamiento y de comparación debe ser el mismo. Por ejemplo, la edad media en el grupo de tratamiento debe ser la misma que la edad media en el grupo de comparación.
- En segundo lugar, los grupos de tratamiento y comparación deben reaccionar al programa de manera similar. Es decir, presentan gustos y respuesta similares.
- En tercer lugar, los grupos de tratamiento y de comparación no pueden estar expuestos a otras intervenciones durante el periodo de evaluación. Por ejemplo, si hemos de aislar el impacto de la transferencia de 20 € sobre el consumo de manzanas, es posible que el grupo de tratamiento también esté recibiendo un bono transporte que le permita ir gratis a la tienda de frutas, mientras que los controles no reciben este bono transporte, por lo tanto, al calcular la diferencia entre el grupo de tratamiento y control se podría confundir los efectos de la transferencia monetaria con el efecto del bono transporte.

Cuando se cumplen estas tres condiciones, entonces solo la existencia del programa de interés explicará cualquier diferencia en el resultado (Y) entre los dos grupos una vez que el programa ha sido implementado. La razón es que la única divergencia entre los grupos de tratamiento y de control es que los miembros del grupo de tratamiento recibirán el programa, mientras que los miembros del grupo de comparación no lo harán. Cuando las diferencias en los resultados pueden atribuirse totalmente al programa, entonces se ha identificado el impacto causal del mismo. Así que en lugar de ver el impacto de la transferencia de renta de 20 € al señor B, el evaluador puede buscar el impacto de un conjunto (muestra) de hombres (que componen el grupo de tratamiento). Si se pudiera identificar a otro grupo de hombres que son totalmente similares al grupo de tratamiento, excepto en el hecho de que estos no reciben la transferencia de 20 €, la estimación del impacto del programa sería la diferencia entre los dos grupos en el consumo promedio de manzanas. Por lo tanto, si el grupo tratado consume un promedio de 6 manzanas por persona, mientras que el grupo de comparación solo consume un promedio de 4, entonces, la fórmula que calcula el impacto del programa es:

$$\alpha = E(Y \mid P = 1) - E(Y \mid P = 0) = 6\text{-}4 = 2 \qquad (2)$$

Donde E() es la esperanza poblacional, utilizando habitualmente su análogo muestral de la media. Por lo tanto, el impacto promedio de la política pública de una transferencia monetaria de 20 € en el consumo de manzanas es de 2 unidades.

Ahora que hemos definido un grupo de comparación válido, es importante considerar lo que pasaría si decidimos seguir adelante con una evaluación sin identificar correctamente este grupo de control, situación que se produce cuando el grupo de control empleado difiere del grupo de tratamiento de alguna manera que no sea la debida a recibir o no la política. Esas diferencias adicionales pueden hacer que nuestra estimación de impacto no sea válida o, en términos estadísticos, sea sesgada: no va a estimar el verdadero impacto del programa. Más bien, se va a estimar el efecto del programa mezclado con el efecto de esas otras diferencias.

$$E(Y \mid P = 1) - E(Y \mid P = 0) = \alpha + \text{sesgo de selección}$$

Es decir, nuestro cálculo de diferencia de la Y entre los dos grupos ya no coincide con el efecto verdadero de la política, sino que nuestra estimación es igual al impacto verdadero (α) más un término adicional, que denominamos sesgo de selección y que analizaremos con posterioridad.

5.3. Métodos para el cálculo del contrafactual

Como se ha mencionado previamente, la pregunta clave que trata de contestar la evaluación de impacto es medir hasta qué punto una determinada intervención pública sobre un conjunto de individuos modifica la variable de resultado de interés, como puede ser el nivel de ingresos, en comparación al valor de esa variable de resultado que los mismos individuos habrían tenido en el caso de que dicha política no hubiera existido, el denominado contrafactual, que es una situación que, por su propia definición, resulta inobservable para el grupo de individuos que reciben el programa.

Así pues, 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, de tal modo que su variable de resultados pueda considerarse equivalente al que habríamos observado para los beneficiarios de la política si esta no les hubiera sido aplicada. Existen dos grandes aproximaciones en la evaluación de impacto para definir este grupo de control. Estos dos métodos difieren entre sí en función del procedimiento utilizado para definir el grupo de individuos que actúan como contrafactual:

- Los diseños experimentales son aquellos en los que, partiendo de una población de potenciales beneficiarios de la 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.
- El resto de métodos disponibles, denominados 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 deciden si participar o no, o debido a que 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.

5.4. El sesgo de selección

Para poder estimar correctamente el efecto de una política, el grupo de comparación debe ser idéntico a los beneficiarios (el grupo de tratamiento) en todos los aspectos excepto en que no reciben la intervención. Sin embargo, la forma en que se seleccionan los beneficiarios (y los no participantes) puede reducir el nivel de comparabilidad entre los grupos de tratamiento y de comparación. A la hora de estudiar los efectos de la política se suele utilizar uno de los dos procedimientos para la selección de los participantes:

- La autoselección, situación en la que se ofrece una determinada política, y en la que las personas están invitados a solicitar, por ejemplo, préstamos para pequeñas empresas, recibir un curso, un tipo de ayuda o que los ayuntamientos deseen participar en un programa de canalización de agua, creación de escuelas y otros servicios sociales, siendo una decisión voluntaria del individuo si participa o no en el programa.
- Selección administrativa, en la que el organismo de ejecución del proyecto selecciona a los individuos, las comunidades o áreas administrativas que van a participar.

Por lo tanto, los participantes pueden tener características especiales, a menudo (cor)relacionadas tanto con la participación o el éxito del proyecto, que los distinguen de los no participantes. En términos econométricos, este es un problema de endogeneidad que sesga las estimaciones de impacto.

Vamos a considerar una política que está focalizada en mujeres entre 30 y 40 años. El objetivo de la política es tratar de aumentar el grado de empleabilidad de este grupo de mujeres y para ello ofrece realizar un curso de capacitación que incrementará sus capacidades en el mercado laboral. Para aumentar el grado de participación, el gobierno ofrece 200 euros mensuales a aquellas mujeres que acudan al menos al 80% de las clases. ¿Qué puede ocurrir? Es posible que las mujeres que componen el grupo de tratamiento y control sean muy distintas. Es más que probable que la mayoría de participantes sean mujeres sin hijos pequeños, ya que su salario de "reserva"

es muy bajo y por lo tanto les resulta interesante participar en el curso de capacitación. Sin embargo, mujeres con niños pequeños, creen que esos 200 euros no compensa el beneficio que les genera cuidar de sus propios hijos y, por lo tanto, rechazarán participar en el programa, así que al final en el grupo de participación puede haber un 80% de mujeres "sin hijos" mientras que en el de control solo exista un 15% de este tipo de mujeres. Este es un ejemplo claro en el que una variable (tener niños pequeños) afecta a la participación en el programa y supone un "sesgo de selección".

Desde un punto de vista analítico, partiendo de la ecuación que calcula la evaluación de impacto de una política:

$$\alpha = E(Y \mid P = 1) - E(Y \mid P = 0)$$

El gran problema en la evaluación radica en que el término $E(Y \mid P = 0)$ es no es observado. La solución al problema depende en parte del tipo de datos disponibles. Los diseños experimentales (también llamados experimentos sociales) utilizan participantes elegibles que han sido excluidos del grupo de tratamiento como indicador de esa situación hipotética. Además, existe otro tipo de estudios (estudios observacionales o cuasiexperimentales) que generan un grupo de comparación utilizando la misma fuente que el grupo tratado, o utilizando otras bases de datos, y, esencialmente, acabando por utilizar una función de $E(Y \mid P = 0)$ que puede ser estimada a partir de datos de los no participantes. La simplicidad a la hora de calcular el estimador en el caso de que los datos provengan de un experimento social (diseño experimental) bien diseñado y ejecutado es una situación completamente opuesta a la que el investigador tiene en la realidad en el contexto de los diseños cuasiexperimental (experimentos observacionales), que están sujetas a un mayor tipo de problemas, como el sesgo de selección o el sesgo de sustitución.

Como ya se ha mencionado, desde un punto de vista estadístico y econométrico, el sesgo de selección surge cuando la variable de tratamiento está correlacionada con el error de la ecuación de la variable de resultado. Esta correlación puede ser debida a que erróneamente se han omitido de variables explicativas observadas que afectan tanto a P como a Y. Entonces el componente omitido que se encuentra dentro del término de error (u) de la ecuación estará correlacionado con P —caso que se conoce como selección en variable observada—. Otra posibilidad es que el sesgo de selección se deba a algún factor no observado que parcialmente determina P además de Y —situación de selección en variable no observada—.

5.4.1. Selección en observables

En muchos diseños cuasiexperimentales en los que el problema de selección se debe a variables observadas, este se puede resolver utilizando métodos de regresión, incluyendo dichas características observadas como variables explicativas del comportamiento de la variable de resultado que estamos analizando. Otra posibilidad para corregir este problema de selección en observables es emplear técnicas de emparejamiento.

5.4.2. Selección en no observables

Sin embargo, existen situaciones en las que las características de selección de los individuos que afectan tanto a la participación como al resultado de la variable de interés no se pueden observar —como puede ser la "inteligencia" de un trabajador, el nivel de "emprendedor" y "sensibilidad por

la tecnología" de un empresario o el espíritu de una "comunidad"—, entonces el uso de las estimaciones que utilizan técnicas de regresión no va a ser capaz de solucionar el problema de sesgo obteniéndose resultados sesgados del impacto del programa público. Sin embargo, en los casos en los que estas características no observadas son invariantes a lo largo del tiempo, se puede cancelar su influencia en el estimador del impacto empleando un método de dobles diferencias, que elimina el sesgo de selección.

5.5. Tipos de evaluación cuantitativa

Dentro de las técnicas de evaluación cuantitativa existen varios tipos de diseño claramente diferenciados que permiten estudiar el efecto en la variable de resultado de largo plazo de un determinado programa. Con todas las posibilidades de evaluaciones, es necesario establecer un procedimiento que permita ayudar a los usuarios a determinar qué tipo de evaluación cuantitativa es mejor para un programa o intervención pública, analizando si el diseño de impacto es apropiado, factible y si es asequible (en términos de costo). La fortaleza de este enfoque radica en que se tiene en cuenta una gama muy completa de tipos de evaluación cuantitativa, y si para cada uno de los programas públicos se examina la conveniencia, viabilidad y asequibilidad de cada uno de estos tipos de diseño, entonces es posible determinar qué tipo, entre todos ellos, es el que presenta mejores propiedades para llevar a cabo una evaluación de ese programa en particular (teniendo en cuenta que esta evaluación sea apropiada, viable y asequible). Los distintos diseños posibles dentro de la evaluación cuantitativa son:

1. Diseño experimental. experimento aleatorio.
2. Diseño de regresión en discontinuidad.
3. Diseño de construcción de grupos de control para emparejamiento.
4. Diseño de series temporales.
5. Diseños de identificación y eliminación causal exhaustiva.
6. Diseño de opinión de expertos.
7. Diseño de opinión de informantes clave.

Como se muestra en la figura de la página siguiente, que establece un diagrama que permite diferenciar entre los distintos enfoques enumerados previamente, se puede establecer una primera diferencia entre los diseños de evaluación cuantitativa dependiendo de si se dispone de contrafactual para realizar el estudio. Por un lado, estarían las aproximaciones (5), (6) y (7), que no emplean grupo de control, mientras que los diseños del (1) al (3) necesitan disponer de un grupo de comparación para evaluar el impacto del programa.

Dentro de los diseños de evaluación de impacto, la primera división se establece entre si se trata de un diseño experimental o no, como ya se vio en la sección anterior. Para cada uno de los diseños, y dependiendo del tipo de sesgo que exista, se determina la técnica de estimación de impacto óptima.

Por el momento estos diseños de evaluación se van a especificar utilizando un lenguaje sencillo empleado en la evaluación y análisis de políticas, aunque es posible realizar una especificación en un lenguaje más técnico y matemático mediante la descripción de estos diseños en forma econométrica o estadística. La tabla 1 muestra los términos técnicos utilizados a la hora de analizar los distintos diseños en la evaluación cuantitativa.

Figura 2. Diagrama de las diferentes evaluaciones cuantitativas

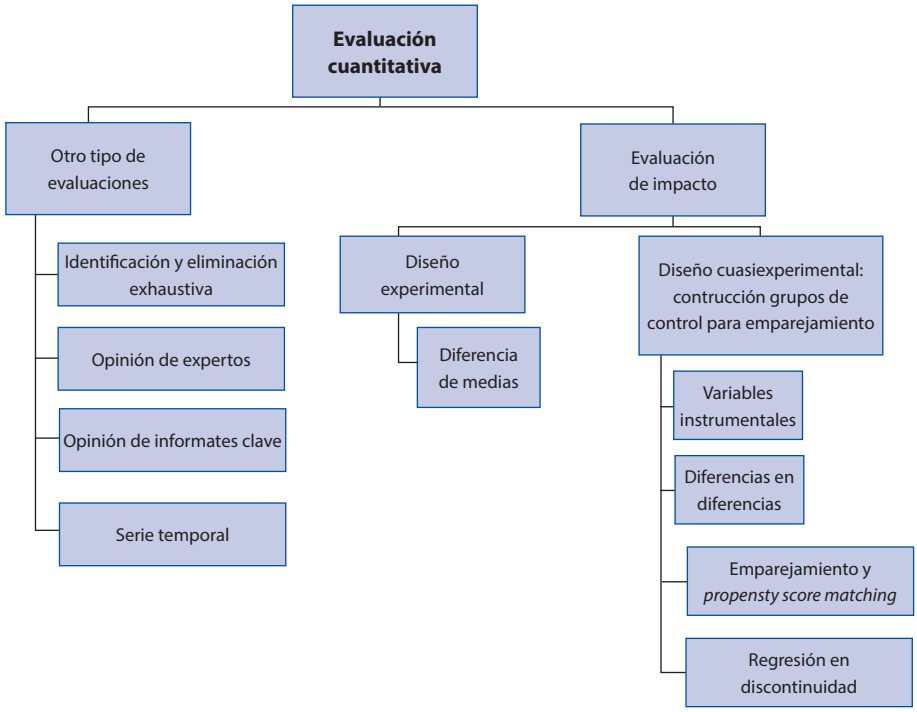


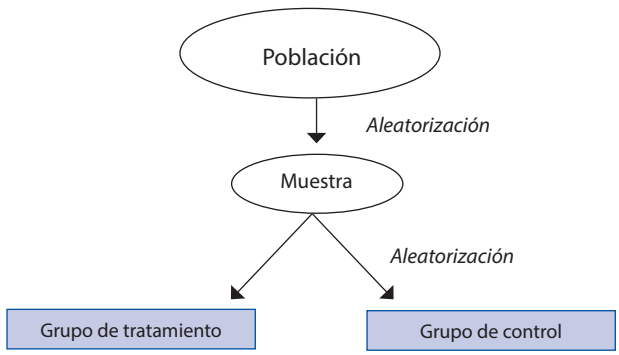
Tabla 1. Nomenclatura de la evaluación cuantitativa

Términos utilizados en los diseños de evaluación cuantitativa	
Unidades	Grupo de individuos (personas, empresas, distritos, etc.) en los que se medirá la variable de resultado de largo plazo (variable de interés o variable de impacto)
Características de unidades	Particularidades de los individuos estudiados (edad, tamaño de la empresa, sexo, población del municipio, etc.)
Modelo de resultados	Diseño del "marco lógico"
Medición	Cuantificación de la variable de resultado de l/p
Grupo de tratamiento	Conjunto de individuos que reciben la intervención pública
Grupo de control	Conjunto de individuos que no reciben la ayuda pública
Intervención	Iniciativa pública que afecta a los individuos analizados. En el marco lógico se debe indicar dónde se produce la intervención
Línea de base o medida preintervención	Cuantificación de la variable de interés antes de que tenga lugar la intervención
Medida posintervención	Cuantificación de la variable de interés en un momento posterior a que finalice la intervención pública

5.6. Diseño experimental

En un diseño experimental puro se dispone de un conjunto de individuos (ya sean estos personas, organizaciones, regiones u otras unidades), que se asignan de manera aleatoria al grupo de tratamiento o al grupo de control, para, a continuación, comparar los cambios en la variable de interés (resultado de largo plazo). Si se observa un efecto en la variable de resultado en el grupo de tratamiento, en comparación al de control, se supone que esta diferencia ha sido causada por la intervención, y no por otro factor, ya que la asignación de las unidades al grupo de control y tratamiento ha sido aleatoria. Existe otra alternativa al diseño experimental puro que es el diseño experimental con lista de espera.

Figura 3. Diagrama de diseño experimental



5.6.1. Diseño experimental con lista de espera

Una variación en el diseño experimental básico es el diseño con "lista de espera". En este diseño las personas (u otras unidades) que desean participar en el experimento y recibir el tratamiento (es decir, están en lista de espera) son asignados de manera aleatoria a recibir la intervención inmediata (grupo de tratamiento), o continuar en lista de espera y recibir la intervención con posterioridad (grupo de control). Sin duda, esta situación no se puede considerar como un diseño experimental puro, situación en la que el grupo de control nunca recibe la intervención. Sin embargo, este diseño se considera a menudo como más adecuado, porque es más ético ya que el grupo de control acaba recibiendo el tratamiento, y más factible, porque los participantes y las partes interesadas son más propensos a aceptarlo, que en el caso de diseños experimentales puros. Sin embargo, el problema con este tipo de diseño es que es necesario poder medir de manera efectiva y real el impacto que la intervención va a ejercer a lo largo del tiempo entre el grupo de tratamiento (que recibe la intervención) y el grupo de control (que la recibirá con posterioridad). Es posible que en el caso de intervenciones en las que sea necesario un tiempo relativamente largo para mejorar los resultados, como ocurre en las políticas educativas, el diseño experimental de lista de espera no resulta el más apropiado.

En el caso de disponer de un diseño experimental, la técnica de estimación de impacto de la política es bastante sencilla, ya que solo es necesario realizar un contraste de medias (entre grupo de tratamiento y control), pudiéndose emplear enfoques paramétricos (contraste de "t") o enfoques no paramétricos, como el test de Kruskal-Wallis.

Existe gran cantidad de trabajos que emplean este tipo de enfoque, a destacar entre otros los de Angrist *et al.* (2002), Banerjee *et al.* (2002) y Behrman y Hoddinott (2005), Duflo *et al.* (2008) y Moffit (2003).

5.6.2. Supuestos del diseño experimental

Los supuestos necesarios para que la aplicación de esta metodología presente buenas propiedades son:

- La media de la variable resultado para el grupo de control es igual a la que hubiera tenido el grupo de tratamiento si no hubiese participado en la intervención.
- La muestra está equilibrada en variables observables y en variables no observadas.

Existen dos situaciones que se producen habitualmente en las que la asignación aleatoria resulta ser un método de evaluación de impacto bastante factible:

1. Cuando la población elegible es mayor que el número de plazas disponibles en el programa. Cuando la demanda de un programa excede el suministro, se puede emplear una simple lotería para seleccionar el grupo de tratamiento dentro de la población elegible. En este contexto, cada unidad de la población tiene la misma oportunidad de ser seleccionada para el programa. Los individuos que son seleccionados en la lotería pasan a formar el grupo de tratamiento, y el resto de la población que no se ofrece el programa pasan a componer el grupo de comparación. En el caso de que existan recursos limitados que impidan la ampliación del programa a toda la población, se pueden mantener los grupos de comparación para medir los efectos a corto, medio y largo plazo del programa. En este contexto, surge el dilema ético de utilizar un grupo de control indefinidamente, ya que un subconjunto de la población necesariamente se quedará fuera del programa.
2. Cuando un programa tiene que implementarse gradualmente hasta cubrir toda la población elegible. Cuando un programa se introducirá gradualmente, el azar determinará el orden en el cual los participantes reciben el programa, dando a cada unidad elegible las mismas posibilidades de recibir tratamiento en la primera fase o en una fase posterior del programa. El grupo que recibe en último lugar sirve como grupo de comparación válido (nuestra estimación del contrafactual). Por ejemplo, supongamos que el Ministerio de Sanidad quiere formar a todos los 15.000 enfermeros en el país para utilizar un nuevo protocolo de la salud, pero necesita tres años para entrenar a todos. En el contexto de una evaluación de impacto, el Ministerio podría seleccionar al azar a un tercio de los enfermeros para recibir formación en el primer año, un tercio para recibir entrenamiento en el segundo año y un tercio para recibir capacitación en el tercer año. Para evaluar el efecto del programa de capacitación de un año después de su puesta en práctica, el grupo de enfermeros formados en el primer año constituiría el grupo de tratamiento y el grupo de enfermeros asignados al azar a la formación en el tercer año sería el grupo de comparación, ya que aún no han recibido la formación.

5.6.3. Fortalezas y debilidades del método

Sin duda, este diseño es el ideal entre todos los posibles enfoques. En el caso de poder implementar un diseño experimental, se van a generar las estimaciones de impacto óptimas,

estadísticamente superiores a cualquier otro tipo de aproximación. Sin embargo, este enfoque presenta ciertas debilidades, que podemos separar en:

¿Son las inferencias válidas para el grupo analizado?

- Fallos en el mecanismo de asignación aleatorio.
- Fallos en la aplicación del protocolo de tratamiento.
- Abandono del programa por parte de los participantes.
- Muestras pequeñas.

¿Podemos generalizar las conclusiones?

- No representatividad de las muestras.
- No representatividad del programa o política.
- La participación es voluntaria.
- Altos costes.
- En algunos casos, problemas éticos.

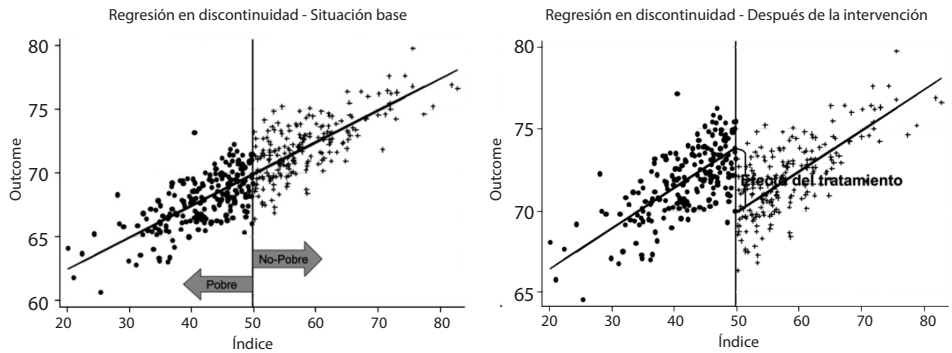
5.7. Diseño de regresión en discontinuidad

El diseño de regresión en discontinuidad emplea otra forma de determinar quién va a ir en el grupo de tratamiento y quién al de control. En estos diseños, las personas, empresas, departamentos u otras unidades se clasifican y se representan gráficamente de acuerdo a una determinada variable que es continua. Por ejemplo, se pueden clasificar a las familias en función de un índice de pobreza[5] para determinar, a partir de un umbral, qué hogares son merecedores de la intervención pública (como, por ejemplo, una actuación encaminada a realizar "transferencias monetarias" o "ayudas alimentarias") y cuáles no. Solo a aquellos hogares por debajo del punto de corte (o valor frontera)[6] en la gráfica se les va a aplicar la política pública (son los elegibles para la intervención). Después de que la intervención ha tenido lugar, y en el caso de que esta hubiera tenido éxito, se debería observar una clara mejora en aquellas unidades que recibieron la intervención, en comparación a aquellos individuos que no la recibieron, es decir, los hogares que se encuentran justo "por encima" del punto de corte. Este diseño cuenta como gran ventaja el que a menudo es visto por los interesados como más aceptable desde un punto de vista ético, ya que, si los recursos son limitados para llevar a cabo la intervención pública, las unidades que reciben la intervención son aquellas que tienen la mayor necesidad de ella.

[5] Otra posibilidad de variable empleada para realizar el corte entre tratados y no tratados puede ser "la nota obtenida en matemáticas o lectura de los estudiantes" o las "tasas de delincuencia" en determinados distritos policiales. "Número de trabajadores" para una política en PYMES, o "edad" de los desempleados en una política de mercado laboral, etc.

[6] Esta variable que indica qué individuos son elegibles para la intervención no es la variable de interés. Es una variable auxiliar.

Figura 4. Regresión en discontinuidad



Dentro de esta metodología de evaluación de impacto se puede consultar los trabajos de Buddelmeyer y Skoufi (2004), Galasso y Ravallion (2004) y Hahn *et al.* (2001).

5.7.1. Supuestos necesarios en el diseño de regresión en discontinuidad

El método tiene que cumplir ciertas condiciones para que sus estimaciones presenten buenas propiedades en términos estadísticos. Primeramente, la selección debe ser determinada por la posición respecto al umbral, definido a lo largo de una variable continua, situación que, por ejemplo, es habitual en las reglas administrativas: variables como el ingreso de hogares y el tamaño de las empresas, la nota obtenida en un examen, medidas de duración o tiempo acumulado en un determinado estado, como desempleo, etc.

Otra condición sobre la aplicabilidad de la RD es que los individuos no pueden ser capaces de manipular su situación respecto al umbral para participar en el programa, conociéndose este problema como la "manipulación de las variables". Supongamos una política que va destinada a PYMES, entonces el número de trabajadores (15) determina si puedes ser beneficiario del programa. Es posible que en esta situación existan empresas que quieran manipular su situación respecto a la frontera (es decir, número de trabajadores que contrata), que claramente afecta a la posibilidad de elegibilidad de la ayuda en sus decisiones de contratación. La ocurrencia de este suceso se puede contrastar porque a la hora de ver la frecuencia de distribución del tamaño de la empresa se observaría un pico en ese valor.

La tercera cuestión que hay que tener en cuenta en el diseño de la RD es la posibilidad de que otro tipo de cambios ocurran en la línea de corte de la variable. Estos cambios pueden afectar a la variable de interés, y este efecto puede ser atribuido erróneamente al tratamiento. Usando el ejemplo previo, supongamos que el valor de 15 trabajadores es el límite para poder aplicar una legislación de protección laboral o para poder ser elegidos para los beneficios de desempleo. Puede ser que ahora, al calcular el impacto, no se pueda separar qué parte de esa cuantía es debida a la nueva legislación y qué parte es debida al programa original de pymes.

5.7.2. Fortalezas y debilidades del método

Como ventaja de este diseño está que permite identificar efectos causales del programa sin imponer restricciones arbitrarias de exclusión, las hipótesis sobre el proceso de selección, las

formas funcionales o supuestos sobre la distribución de los errores. El diseño RD puede ser la mejor alternativa a los estudios aleatorizados para evaluar la efectividad del programa. El elemento más importante del diseño RD es el uso de la puntuación de un "corte" en una medida pretest para determinar la asignación a intervención o control. Una característica importante de esta técnica es que la medida de la selección no tiene por qué ser la misma que la medida de resultado, maximizando así la capacidad del programa para utilizar las guías de práctica basadas en la investigación, instrumentos de encuestas y otras herramientas para identificar a las personas más necesitadas de la intervención del programa. Otras posibles ventajas que merecen ser destacadas son:

- Nos ofrece una estimación insesgada del efecto del tratamiento en la discontinuidad.
- Muchas veces, si se tiene una regla conocida para determinar qué individuos pertenecen al grupo de beneficiarios y cuáles al de control supone una ventaja. Este tipo de reglas son comunes en el diseño de la política social (una definición de pobreza es aquella situación en la que un individuo tiene una renta inferior al 60% de la mediana o media de la distribución de ingresos de la población de referencia).

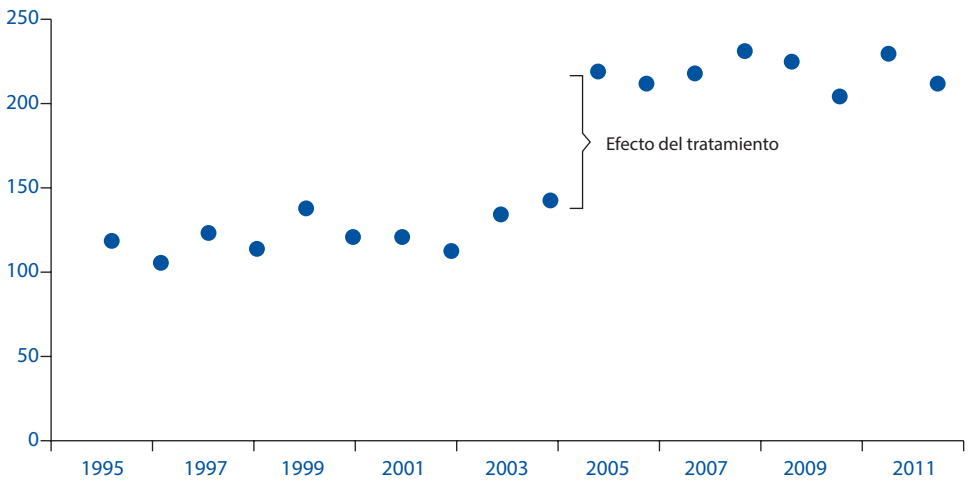
Por otro lado, este diseño presenta dos limitaciones notables. En primer lugar, su viabilidad está, por definición, limitada a aquellos casos en los que la selección se lleva a cabo en una medida previa a la intervención observable, y no suele ser un hecho que se dé habitualmente. En segundo lugar, incluso cuando el diseño es factible, solo identifica el impacto medio en el umbral para la selección. Lo que en presencia de efectos heterogéneos en los individuos no dice nada relevante sobre el impacto en las unidades de distancia del umbral de selección. En este sentido, solo se identifica un impacto medio local del tratamiento. Otras desventajas son:

- Estimación local: los efectos del tratamiento alrededor del corte del índice no siempre son generalizables.
- Potencia: el efecto es estimado en la discontinuidad, así generalmente tenemos menos observaciones que en un experimento aleatorio para un mismo tamaño de muestra.
- La especificación puede ser sensible a la forma funcional: tenemos que modelar correctamente la relación entre la variable de asignación y la variable de resultado.
  - Relaciones no lineales.
  - Interacciones.

5.8. Diseño de series temporales

Un diseño de series temporales utiliza el hecho de poder disponer de una sucesión suficientemente larga y amplia en el tiempo de mediciones de la variable de interés o variable de impacto. En estos casos, una vez que ha pasado algo de tiempo (es decir, tenemos información previa a la intervención), se realiza la política pública, y es necesario saber si el programa ha tenido el efecto buscado. Si esto es así debemos ser capaces de observar un cambio significativo en el nivel de la serie temporal de la variable de interés en un determinado momento de tiempo, que es cuando ocurrió la intervención pública.

Figura 5. Serie temporal



Dentro de las técnicas de series temporales se recomienda consultar el libro de Harvey (1990).

5.8.1. Supuestos necesarios en el diseño de series temporales

La utilización de este diseño con plenas garantías de éxito se basa en los siguientes supuestos:

- En primer lugar, la existencia de una fecha claramente identificable en la que se produce la intervención (ejemplo: en el año "t" se incluye en la declaración de la renta de las personas una deducción por compra de vivienda).
- En segundo lugar, se basa en la existencia de suficientes puntos de datos, de tal forma que sea posible ver que la tendencia preexistente en la variable de resultados, es decir, datos previos a la intervención, se modificó cuando se inicia la intervención.
- Si no hubiese tenido lugar la reforma, la media de la variable resultado sería la misma antes y después de la reforma.
- No hay efecto de anticipación a la reforma.

5.9. Diseño de construcción de un grupo de control para emparejamiento

Se dispone de este tipo de diseño de evaluación cuando es posible encontrar un grupo de individuos que presentan características similares en muchos aspectos al grupo de tratamiento, salvo por el hecho de que estos individuos no están recibiendo la intervención. Por ejemplo, supongamos que tenemos un programa que se implementa en un(os) distrito(s) de la ciudad. En esta situación, el grupo de control podría estar compuesto por distritos que son similares a los distritos sobre los que se realizó la política, pero que no la han recibido. Otra opción seria la utilización de diferentes ciudades o departamentos.

Existe una variación de esta versión de emparejamiento, que emplea la misma lógica subyacente, y consiste en realizar estimaciones de lo que ocurre en promedio a las personas que presentan una determinada probabilidad de recibir un tratamiento, en función de un

conjunto de características (por ejemplo, en el mercado laboral, la edad, tiempo de desempleo, si están casados, etc.). Supongamos que se tiene una política de empleo que pretende mejorar el nivel de empleabilidad de personas que llevan en situación de desempleo durante más de un año. Entonces se decide realizar una intervención en un grupo y se compara lo que les sucede (el tiempo que permanecen en el subsidio de desempleo) con el tiempo que deberían haber permanecido recibiendo la prestación por desempleo si la intervención no hubiera tenido efecto —utilizando como grupo de comparación individuos que presentan probabilidades (o propensiones) similares de participar en el programa público, pero que en realidad no participaron—. Este subtipo de este diseño se llama *propensity score matching*. Sin embargo, surgen ciertos problemas cuando se construyen grupos de emparejamiento debido a que, en comparación con los experimentos verdaderos, es muy probable que el grupo de tratamiento sea diferente (en algún aspecto, observado o no) respecto al grupo de control, inconveniente que se conoce habitualmente como el problema de "sesgos de selección" en la evaluación de impacto. Para tratar de solucionar este problema, existe una serie de técnicas econométricas y estadísticas. Las más habituales en la literatura son:

- Técnica de diferencias en diferencias.
- Método de variables instrumentales.
- Emparejamiento y *propensity score matching*.

A continuación pasamos a analizar los supuestos necesarios para cada una de las técnicas de evaluación asociadas con la construcción de grupos de control para emparejamiento.

5.10. Diseño de variables instrumentales

Existen multitud de ejemplos en los que la técnica de variables instrumentales se pueden aplicar: cambios en las reglas administrativas entre jurisdicciones limítrofes, cambios súbitos en la legislación, debido a modificaciones de reglas políticas, factores geográficos, como un cambio en la proximidad entre el cliente y el suministrador del servicio, descensos inesperados del presupuesto de un programa, cambios en las "condiciones administrativas". Todos estos ejemplos producen experimentos naturales. Estas "fuerzas externas", en términos econométricos, se denominan "instrumentos". Más aún, los métodos de VI son aplicables a todas aquellas situaciones en las que el acceso al programa está sujeto a la aleatorización, pero los agentes afectados (clientes o suministradores) no están completamente de acuerdo, por lo que se genera una situación en la que el acceso al programa está determinado tanto por las preferencias de los individuos como por la aleatorización. Finalmente, se puede considerar una situación en la que la estimación del impacto de la política se puede obtener mediante un mecanismo: promover que algunos individuos y no otros han sido seleccionados en dos grupos, de manera aleatoria, a formar parte del programa.

La siguiente tabla muestra las diferentes situaciones en las que se aplica el método de variables instrumentales.

Tabla 2. Situaciones en las que se aplica el método de VI

Aleatorización y perfecta aceptación	Aleatorización con aceptación parcial	Promoción aleatoria	Experimento natural no aleatorio
Existe un elemento de manipulación por parte del investigador			No existe manipulación deliberada
Utilizar diferencia de medias	Utilizar método de variables instrumentales		
El efecto del tratamiento se identifica mediante la diferencia entre las medias de la variable de interés entre tratamiento y no tratamiento	Estimador de Wald: El efecto del tratamiento se identifica mediante el cociente de dos estimaciones: en el numerador la diferencia de medias entre elegibles y no elegibles. En el denominador la probabilidad de tratamiento inducida por el instrumento		Estimador de dos etapas: 1ª etapa: el modelo estima la probabilidad de tratamiento como función del instrumento y más variables. 2ª etapa: se estima la ecuación de la variable resultado usando la predicción de la probabilidad de tratamiento

Diferentes trabajos que utilizan este método son los realizados por Abadie *et al.* (2002), Heckman y Vytlacil (2000), Angrist (1990), Blundell y Costas-Dias (2008) y Heckman (1997).

5.10.1. Supuestos necesarios en el diseño de variables instrumentales

Para la correcta aplicación de este método debe existir una variable auxiliar, también llamada "instrumento", y que denominamos "Z" que debe cumplir las siguientes propiedades simultáneamente:

- La variable "Z" está altamente relacionada con la variable que desea instrumentar. Ejemplo, si creemos que la variable explicativa "ir a la universidad" tiene problemas de endogeneidad, un posible instrumento es "si la universidad está cerca o no".
- La variable "Z" no se encuentra relacionada con el término de error de la ecuación de regresión que analiza el efecto de la política en la variable de interés.
- La variable "Z" debe afectar a la participación en la política, pero no la variable de interés "Y".

5.10.2. Fortalezas y debilidades del método

La mayor debilidad de este método es que puede ser difícil encontrar un instrumento que sea a la vez relevante y exógeno. La evaluación de la exogeneidad del instrumento puede ser algo subjetivo. Aún más, el método de VI resulta difícil de explicar para aquellos que no están familiarizados con ella.

La mayor fortaleza es el hecho de explotar situaciones que son similares a la experimentación aleatorizada. Además, las situaciones en las que se usan instrumentos generados por el propio investigador han aumentado, lo que refleja la convergencia entre la experimentación clásica y los métodos de investigación observacionales. El desarrollo más importante es el uso de las

variables instrumentales en los experimentos de aleatorización. Las variables instrumentales son útiles en experimentos cuando, tanto por consideraciones prácticas como éticas, existe una conformidad incompleta en los grupos de tratamiento y de control. En la evaluación aleatorizada de los programas de capacitación, por ejemplo, algunos miembros del grupo de tratamiento pueden rechazar el curso mientras que algunos del grupo de control aprovechan el curso mediante canales fuera del experimento.

Como en los experimentos naturales, el instrumento suele explotar un origen exógeno de variación —creado mediante una asignación explícita aleatoria en estos casos— para estimar el efecto de interés. Similarmente, en los ensayos médicos, los doctores pueden intentar aleatoriamente ofrecer, y no imponer, incentivos para el cambio de comportamientos como fumar o tomar un medicamento.

El progreso en la aplicación del método de variables instrumentales depende considerablemente del trabajo desarrollado en encontrar y crear un instrumento creíble que se puede usar para medir las relaciones económicas importantes. Aquí los retos no son tanto en términos técnicos en el sentido de que se requieren nuevos teoremas o estimadores. La clave está en el conocimiento detallado de la situación institucional y en una investigación cuidadosa y cuantificación de las fuerzas en el trabajo en una determinada situación particular.

5.11. Diseño de emparejamiento o 'propensity score matching'

La idea que se esconde detrás del *matching* consiste simplemente en seleccionar un grupo de no beneficiarios con el fin de hacerles lo más parecidos a los beneficiarios en todo, salvo en el hecho de haber recibido la intervención, si se logra hacer que este grupo sea lo más parecido, entonces las variables de interés observadas en el grupo emparejado se aproxima al contrafactual, y el efecto de la intervención se estima como la diferencia entre las medias de las variable de resultado de los dos grupos. Por ejemplo, para estimar el efecto de las ayudas en el incremento del gasto en I+D, se emparejan empresas ayudadas con un conjunto de empresas no ayudadas que se parecen en todas las variables explicativas relacionadas con el proceso de participación en el programa. El efecto de la ayuda en el gasto de I+D es estimado mediante la diferencia entre la media del gasto de I+D de las empresas ayudadas menos el gasto medio de las no ayudadas emparejadas. Todo esto bajo la condición de que el emparejamiento genera dos grupos equivalentes.

La técnica de *matching* exacto trata de emparejar individuos tratados con no tratados en función de sus características observadas. El caso del "emparejamiento exacto" se puede realizar cuando el vector de variables discretas observadas es discreto y la muestra contiene suficientes observaciones para cada uno de los distintos valores de Xi. Encontrar un buen emparejamiento para cada participante del programa requiere aproximar lo más posible las variables o determinantes que explican la decisión de que la persona que se inscriba en el programa. Desafortunadamente, esto es más fácil decirlo que hacerlo. En el caso de:

1. Si la lista de características observadas pertinentes es muy grande.
2. Si cada característica adquiere muchos valores o es una variable continua, como la edad o el salario.
3. Si la muestra de datos es pequeña.

En estas situaciones puede resultar bastante difícil identificar algún individuo para cada una de las unidades en el grupo de tratamiento. A medida que aumenta el número de características o dimensiones para hacer coincidir las unidades que participan en el programa, es posible que aparezca lo que se conoce como "la maldición de la dimensionalidad".

La solución a este problema consiste en utilizar como variables de emparejamiento no a las X, sino a una nueva variable que es la probabilidad que tiene un individuo de participar en el programa en función de las variables explicativas X. Esta nueva aproximación basada en una probabilidad, también llamada *propensity score* se debe a Rosenbaum y Rubin (1983), que permite convertir un problema multidimensional en un problema unidimensional, evitando de ese modo el problema de maldición de la dimensionalidad.

El *propensity score* se define como la probabilidad condicional de recibir tratamiento en función de una serie de variables observadas X antes de tratamiento, que se expresa como:

$$p(X) = Pr(D = 1 \mid X)$$

Donde P(x) es la probabilidad de que ocurra el suceso que aparece dentro del paréntesis. Esta probabilidad se utilizará para calcular en la segunda etapa el efecto de la politica. Trabajos que emplean esta técnica de evaluacion se encuentran en Bryson *et al.* (2002), Caliendo y Koepeining (2008), Dehejia (2005), Hahn *et al.* (2008) y Heckman *et al.* (1995).

5.11.1. Supuestos necesarios en emparejamiento o 'matching'

Los dos supuestos fundamentales bajo los cuales este diseño presenta buenas propiedades en la estimación del efecto de una política son:

- Supuesto de independencia condicionada: no existen variables no observables que influyan, a su vez, en la participación en la intervención y en el resultado.
- La intersección de los valores del *propensity score* para los grupos de tratamiento y control no es vacía, por lo que existe un soporte común para los dos grupos.

Pasamos a continuación a desarrollar cada uno de ellos.

5.11.1.1. Supuesto de independencia condicionada

Una condición esencial para la aplicabilidad de este método consiste en poder disponer de las características de los individuos antes de que la intervención se lleve a cabo. En el caso de usar variables observadas después de la intervención se está cometiendo un error, ya que es posible que estas puedan ser influenciadas por la intervención. Idealmente, todas las variables que afectan al proceso de selección deben estar incluidas en la lista de variables del *matching*, aunque esta situación no suele ser la más habitual.

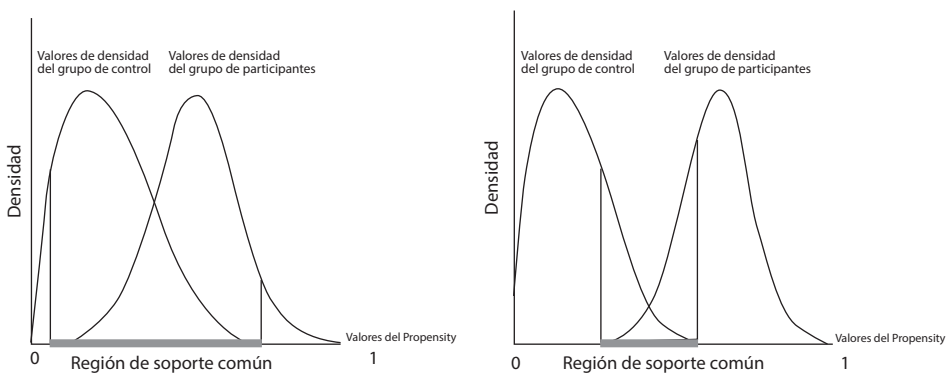
El supuesto de independencia condicional consiste en que dado un conjunto de variables X que no están afectadas por el tratamiento, los resultados potenciales Y son independientes de la asignación del tratamiento P. Este supuesto también se conoce como "ausencia de factores de confusión", e implica que realizar o no un programa depende exclusivamente de las características observadas.

En el caso de disponer de un conjunto de datos previos a la realización del programa es posible que se den las condiciones necesarias para que este supuesto de independencia condicionada se cumpla, ya que permite al investigador controlar aquellas características observadas que parecen afectar a la participación en el programa (suponiendo que la selección no observada es bastante pequeña). En el caso de que existan variables no observadas que afectan a la participación existen métodos alternativos de estimación del impacto del programa como son las variables instrumentales o el método de DD.

5.11.1.2. Supuesto de soporte común

Otra condición necesaria para la correcta aplicación de esta técnica de emparejamiento es la existencia de un solapamiento considerable entre las tasas de participación de los beneficiarios y no beneficiarios. Esta superposición se conoce como el "soporte común". La representación más intuitiva del problema de soporte común se da en la figura 6: el área entre las dos barras verticales, en la que se pueden encontrar los beneficiarios y no beneficiarios que comparten valores similares para la probabilidad de estar expuestos a la intervención, es el soporte común. Aquellos beneficiarios que con un valor muy bajo en la probabilidad (zona a la izquierda de la primera línea vertical), y los beneficiarios con valores de propensión muy alta a la derecha de la segunda línea vertical, deberían excluirse del análisis para así realizar una comparación correcta.

**Figura 6. Ejemplo de soporte común compacto (izda.)
y problemas de soporte común compacto (dcha.)**



Por lo tanto, las unidades del grupo de control deben ser semejantes a las del grupo de tratamiento en las variables observadas que no han sido afectadas por la intervención, así que será necesario eliminar aquellas observaciones del grupo de control que no permiten garantizar la correcta comparabilidad de los dos grupos. Del mismo modo, a veces es necesario borrar un conjunto de datos de aquellas unidades del grupo de tratamiento para el que no se localiza ningún individuo del grupo de control semejante. Esta situación es bastante complicada ya que puede generar un sesgo en el efecto del tratamiento, por lo que será necesario interpretar el sesgo potencial en la estimación de los efectos de tratamiento.

5.11.2. Fortalezas y debilidades del método

El *matching* tiene dos claras desventajas en relación con el diseño experimental (que utiliza el contraste de medias para estimar el impacto de la política). La primera es la necesidad de asumir la independencia-condicional, que permite eliminar el sesgo de selección mediante el control en las variables observables. En el caso de que la asignación aleatoria se realice correctamente, podemos estar seguros de que las poblaciones beneficiarias y no beneficiarias son similares tanto en las características observables y no observables. En segundo lugar, mientras que la técnica de emparejamiento solo puede estimar los efectos del tratamiento en donde existe una superposición entre los beneficiarios y la población, la asignación aleatoria asegura que existe un soporte común a través de la muestra de no beneficiarios. Estas consideraciones hacen que el diseño experimental sea mejor. Sin embargo, las consideraciones prácticas también son importantes en el diseño y ejecución de las evaluaciones de los programas y, a menudo, al tener en cuenta estas características empíricas favorecen al *matching* sobre la asignación aleatoria.

La principal ventaja del *matching* y PSM sobre el diseño experimental es que evita las consideraciones éticas que surgen cuando un tratamiento potencialmente beneficioso se niega por razones de "azar". El costo también es una consideración práctica importante cuando se realizan las evaluaciones. En algunos casos, a pesar de que los requisitos de datos del diseño del emparejamiento es considerable, la generación de datos puede ser menos costoso que en el caso de un experimento, ya que este último implica una monitorización sustancial para asegurar la asignación al azar.

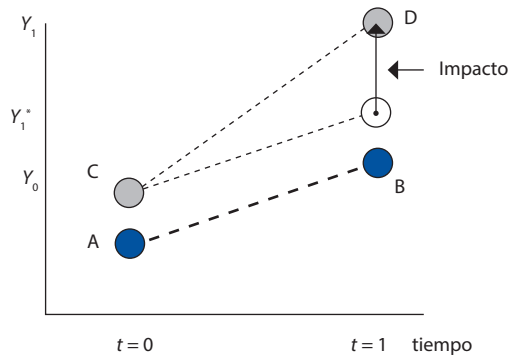
Por lo que se refiere a las diferencias entre el *matching* y las otras técnicas de evaluación no experimental, la técnica de emparejamiento es mejor que los métodos de regresión estándar por dos razones. En primer lugar, los estimadores de *matching* tienen en cuenta el problema de soporte común. Cuando no existe una buena superposición entre los beneficiarios y no beneficiarios se plantean preguntas sobre la solidez de los métodos tradicionales. En segundo lugar, el *matching* no necesita supuestos sobre la forma funcional para la ecuación de resultado. Los métodos de regresión imponen una forma de relaciones (por lo general, lineal) que puede (o no) ser exacta y que la técnica de *matching* evita, lo que es de gran valor ya que estas restricciones en la forma funcional de la regresión no están justificadas ni por la teoría ni los datos utilizados.

Una crítica típica a la técnica del PSM es que el emparejamiento de individuos del grupo de control y tratamiento lo realiza como una "caja negra", sin saber muy bien cómo funciona el programa.

5.12. Diseño de diferencias en diferencias

Desde un punto de vista analítico, el método de diferencias en diferencias consiste en tener información del grupo de control y tratamiento en dos periodos, antes y después de que se produzca la intervención pública. El método consistirá exclusivamente en calcular las diferencias en la evolución temporal de cada grupo (primera diferencia) para con posterioridad ver el diferencial en el crecimiento que se produjo entre el control y tratamiento (segunda diferencia). Desde un punto de vista gráfico es:

Figura 7. Diseño de diferencias en diferencias



Esta técnica se puede estudiar con más detenimiento en Card y Krueger (1994), Bell (1999), Bertrand *et al.* (2004) y Chaudhury y Parajuli (2006),entre otros.

5.12.1. Supuestos necesarios para diseño de diferencias en diferencias

Las condiciones que se deben cumplir para que las estimaciones obtenidas con esta técnica presenten buenas propiedades son:

- El crecimiento de la variable resultado entre antes y después de la reforma para los no participantes es igual que para los participantes si la reforma no hubiera tenido lugar.
- En particular, el grupo de control reacciona a acontecimientos coincidentes con la intervención igual que el de tratamiento.
- La muestra está equilibrada en variables observables.
- No hay efecto de anticipación a la reforma.

La aplicación del método DID necesita que la variable de interés se pueda medir varias veces a lo largo del tiempo, es decir, es posible tomar mediciones equivalentes en distintos momentos de tiempo y estas mediciones se pueden hacer independientemente de la existencia de una determinada acción pública. Incluso existe la posibilidad de replicar el mismo tipo de medición a lo largo del tiempo sobre las mismas unidades (ventas de empresas, ingresos de hogares, salarios de trabajadores). En esta situación estamos trabajando con datos de panel.

Algunas variables de resultado tienen sentido medirlas solo una vez en cada uno de los individuos, como la duración del desempleo una vez que se ha perdido el puesto de trabajo, el peso de un recién nacido... En esta situación, la obtención de resultados creíbles se basa en la obtención de información a nivel más agregado usando cohortes sucesivas de individuos que experimentan el mismo suceso. Por ejemplo, sucesivas generaciones de individuos, que pasan a ser desempleados, crearán diferentes estimaciones de la duración media de desempleo.

Otra característica relevante a tener en cuenta en la posible aplicación del DID es si los datos de la variable de interés se recopilan de manera rutinaria por parte de estadísticas oficiales, como la tasa de empleo o el PIB per cápita, o la posibilidad de que los datos sean coleccionados

ad hoc. En este último caso, el gran problema para aplicar la DID es que habitualmente no existe ningún levantamiento de datos anterior a la realización de la política. Si no hay posibilidad de datos previos a la intervención, existe la opción de obtener información de manera retrospectiva para el periodo antes de la aplicación de la política. El peligro de este tipo de estrategia es la contaminación entre las mediciones de distintos periodos de tiempo pero tomados en la misma entrevista.

La aplicación de esta técnica también requiere que la intervención sea de naturaleza discreta, es necesario que existan ciertos individuos que estén expuestos a la política y otros individuos que no. Las intervenciones que tienen un carácter continuo no se pueden analizar de forma sencilla con este tipo de método.

5.12.2. Fortalezas y limitaciones de la técnica

A pesar de ser uno de los métodos más utilizados, el diseño de DID no es la panacea que soluciona todos los problemas que existen en la estimación de la evaluación de impacto. Como ventajas claras de esta aproximación están el que ya no es necesario tener estructuras de datos muy complejas, y para evaluar solo son necesarios datos agregados, obtenidos antes y después de la intervención. Además es capaz de corregir sesgos debidos a variable no observada —corrección que no son capaces de afrontar otros diseños—, siempre y cuando este sesgo sea constante en el tiempo.

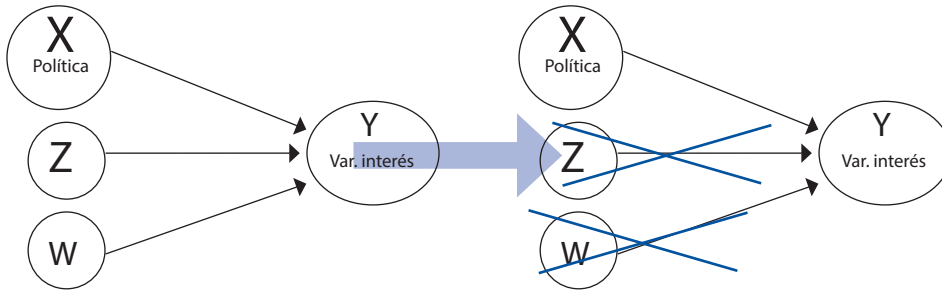
Las limitaciones de esta técnica comienzan cuando se aplica este método en la práctica. El primero de ellos es la necesidad de datos antes de la aplicación del programa público, que suele ser un obstáculo bastante complicado de solventar, ya que existen bastantes lagunas en la planificación de la recolección de datos para evaluar la política. Desde un punto de vista más conceptual, la simplicidad del método paga un precio a la hora de los supuestos necesarios para su aplicabilidad: el supuesto fundamental para la identificación del impacto es que la tendencia temporal del grupo de control y del grupo de tratamiento es similar. La cuestión es que este supuesto se puede contrastar solo si se disponen de datos en más años.

5.13. Diseño de identificación y eliminación causal exhaustiva

El diseño de identificación y eliminación causal exhaustiva funciona en los casos en los que se ha observado una mejora en la variable de interés (impacto), pero no se tiene claro si esta mejora se debe a la intervención o por algún otro factor. Cuando se utiliza este diseño, el primer paso consiste en enumerar todas las posibles explicaciones alternativas (otros factores que afectan a la variable de interés), para, a continuación, tratar de eliminar cada uno de ellos de manera sistemática como causante de la mejora. Si usando esta técnica es posible eliminar todos estos otros factores, entonces, es razonable creer que el programa es el causante de la mejora en la variable de resultados.

Si se desea más información, las obras de Weiss (1992) y Hermatti (2002) pueden resultar de gran interés.

Figura 8. Diseño de eliminación causal exhaustiva



5.14. Diseño de opinión de expertos

En este diseño, se solicita a un experto que realice un juicio sobre si los resultados en la variable de interés (la de impacto) son atribuibles a una intervención pública. Es de esperar que el experto utilice cualquier tipo de datos e información que haya podido recopilar, así como distintos métodos de análisis que normalmente utiliza en su área de trabajo, y de este modo poder aprovechar todos sus conocimientos previos en casos similares[7].

Si se desea conocer más de este método, se recomienda consultar los trabajos de Nadeau M-A. (1988), Witkin y Altschuld (1995), Callon *et al.* (1995) y Cozzens (1987).

5.15. Diseño de juicio de informantes clave

En este diseño, se pide a los informantes clave de la política (personas que tienen experiencia en el programa o en aspectos significativos del programa) que realicen un juicio sobre si, en su opinión, los cambios en la variable de interés son atribuibles a la intervención. Se espera que se utilice cualquier dato o método de recolección y análisis que normalmente emplearan estas personas en su día a día de trabajo y aprovechar sus conocimientos previos en casos similares. Con posterioridad, estos juicios se agregan y resumen para conformar un conjunto de conclusiones sobre los resultados del programa.

Información sobre este tipo de técnica se puede encontrar en Bryk (1983), Healely (1998), ODA (1995) y Weiss (1998).

7 Determinados grupos de interés en algunos casos no aceptan este diseño como una técnica propiamente dicha de evaluación de impacto por no ser lo suficientemente robusta en términos matemáticos y estadísticos.

6. Selección del método de evaluación cuantitativa mediante una lista guía de decisión

Determinar cuál es la técnica de evaluación cuantitativa óptima puede ser una tarea complicada debido al hecho de que hay una amplia gama de opciones en el diseño de estas evaluaciones, como se observa en Weiss (1998) y Duignan (2009). En esta sección vamos a tratar de presentar un conjunto de reglas de decisión, mediante tablas, que permitan ayudar a decidir sobre cuál es el tipo de evaluación de impacto más adecuada a utilizar en determinadas situaciones[8]. Además, para hacer frente al problema de la existencia de grupos de comparación potencialmente no equivalentes, hecho bastante habitual en la práctica de la evaluación, también se va a proponer un procedimiento para seleccionar la técnica óptima de evaluación en los esquemas cuasiexperimentales y, por lo tanto, diseñar grupos de comparación emparejados. Por lo tanto, la estructura secuencial que presenta esta sección es:

1. Etapa 1. Guía de decisión (utilizando la tabla 3) para determinar el método de evaluación cuantitativa más adecuado en función de las características más importantes de la intervención pública a evaluar.
2. Etapa 2. Evaluación de impacto de diseños cuasiexperimentales. Construcción del grupo de control para emparejamiento, debido a la existencia de sesgos de selección, mediante una guía, dada por la tabla 4, que determina cuál de las técnicas de construcción del grupo de control emparejado es el más adecuado.

Para finalizar esta sección se ofrece un árbol de reglas de decisión (figura 9), que indica, en función del tipo de información disponible, qué tipo de evaluación de impacto, dentro de aquellas que utilizan contrafactuales, es la más apropiada.

6.1. Etapa 1. Utilización de una tabla de decisión para determinar el método de evaluación

El método para decidir la mejor técnica de evaluación cuantitativa es en un proceso de dos etapas. En la primera etapa se determina el diseño de evaluación de impacto óptimo en función de las características más importantes de la intervención pública, como son: si se controla quién recibe la intervención, si se aplica la medida política a todos los individuos, si existen recursos limitados para realizar la evaluación, etc. En el caso de que la técnica de evaluación cuantitativa óptima sea el diseño de la construcción de un grupo de control para emparejamiento, entonces se produce una segunda etapa. En esta fase, se ofrece un método de selección para determinar cuál de los cuatro posibles métodos de grupo de control emparejado considerados es más adecuado.

Antes de pasar a analizar en detalle el proceso de elección de la técnica de evaluación es necesario presentar cierta información preliminar.

8 En el anexo II se proporcionan una serie de *checklists* (listas de validación) para cada uno de los posibles diseños de evaluación de impacto. Estas listas de validación se pueden utilizar para decidir cuál es el mejor diseño de evaluación en términos de idoneidad, viabilidad y asequibilidad, y también para revisar las características existentes entre distintos diseños de evaluación de impacto que se han propuesto o que se han llevado a cabo.

6.1.1. Información preliminar

En esta sección se expone la información necesaria para poder analizar el contenido de las tablas, de tal modo que se pueda comprender mejor las opciones contenidas en ellas, y que, por lo tanto, resulte más sencillo poder tomar decisiones para cualquier usuario.

Tipos de evaluación

La tabla 3 que se mostrará para la toma de decisiones tiene en cuenta los siguientes diseños de evaluación de impacto, estas son:

- Diseño experimental.
- Diseño de regresión en discontinuidad.
- Diseño de series temporales.
- Diseño de construcción de grupos de control para emparejamiento.
- Diseño de identificación causal y eliminación exhaustiva.
- Diseño de juicio de expertos.
- Diseño de juicio de agentes clave.

Es necesario recordar que no todos estos diseños (especialmente los dos últimos) son habitualmente considerados por todas las partes interesadas en el seguimiento de políticas públicas como técnicas de evaluación lo suficientemente robustas. Sin embargo, todos ellos son aceptados en algunas situaciones como apropiados, así como factibles y asequibles, y, por lo tanto, se incluyen en aras de una mayor exhaustividad en el estudio que se desea realizar.

6.1.2. Tabla 3. Toma de decisiones para la selección de la técnica de evaluación cuantitativa

La toma de decisiones de la tabla 3 se centra en la lista de los posibles diseños de evaluación de impacto que se pueden realizar, que van desde una óptica más cuantitativa como los "diseños experimentales" hasta los de menor carga numérica como los "juicio de expertos". Esta tabla nos permite, mediante un método asistido, observar las principales características de la intervención pública e identificar aquel diseño de evaluación de impacto con mejores propiedades.

En la tabla 3, las filas representan las distintas técnicas de evaluación y las columnas, las características de la intervención pública. Para usar esta tabla de manera correcta el paso inicial consiste en buscar en la parte superior de la tabla (en la primera fila) e identificar aquel(los) elemento(s) o características que existen en la intervención que se desea evaluar. Cuando se localiza una de estas características, el siguiente paso consiste en mirar en toda esa columna situada inmediatamente debajo de la celda seleccionada, lo que nos indica qué apropiado, viable o costoso resulta cada una de las posibles evaluaciones de impacto aquí planteadas. A continuación, una vez seleccionada aquella celda que presenta mejores propiedades (es decir, que sea viable, idónea o asequible), se observa la parte de la izquierda de la tabla (la primera columna) y se ve cuál de los posibles tipos de evaluación de impacto puede ser el más apropiado en esa situación en particular. Para ello, vamos a considerar que una intervención puede presentar una o más de las siguientes características (fila superior de la tabla 3):

- No se puede controlar quién recibe la intervención. Cuando los evaluadores no pueden controlar quién recibe una intervención, esto limita seriamente la gama de los posibles diseños de evaluación de impacto que se pueden utilizar.
- No se puede evitar que el grupo de control pueda recibir la intervención. En algunos casos, los evaluadores no pueden evitar el hecho de que un grupo de control o comparación pueda obtener alguna versión de la intervención. Obviamente, esto causa problemas a cualquier diseño de evaluación basado en la comparación de los resultados entre un grupo de tratamiento respecto a los resultados del grupo de control.
- Todos los individuos reciben la intervención. Cuando se implementa un programa público a todo el mundo (todas las personas o unidades) Esto limita gravemente la variedad de diseños de evaluación de impacto que se pueden utilizar, debido a la dificultad de poder encontrar un grupo de control.
- Recursos limitados para la realización de una evaluación de impacto. A menudo hay recursos limitados para evaluar un programa, y como consecuencia, para poder realizar evaluaciones de impacto. Esto puede influir decisivamente en qué método de evaluación se puede emplear para estudiar un programa público.
- Preocupación porque no se complete la evaluación de impacto. Es posible malgastar importantes recursos en evaluaciones de impacto que nunca se llegan a finalizar, porque los planificadores de la evaluación subestiman la aparición de problemas prácticos que impidan completar hasta el final la evaluación. Es necesario que en el diseño de la evaluación se tenga en cuenta este tipo de riesgo.
- Actores importantes, sobre todo responsables políticos, son escépticos acerca de la intervención política. Cuando los actores relevantes en la política son escépticos acerca de la efectividad de esta intervención, es necesario considerar aquellas formas más intensas en términos cuantitativos de evaluación para que estas partes interesadas vean cómo se "demuestra" que una determinada acción pública ha mejorado los resultados de la variable de interés.
- Dificultades, para ciertos informantes clave o expertos, en la determinación de las relaciones causales de la intervención. Debido a la naturaleza de algunas intervenciones, resulta relativamente fácil observar la relación de causalidad entre el programa público y la variable de resultados. Sin embargo, en otros casos, es posible que existan muchos factores que están influyendo en los resultados finales. En esta situación, es posible que distintos diseños sean los más apropiados dependiendo de las circunstancias que se tienen en cuenta.

Tabla 3. Selección del método de evaluación en función de las características de la intervención pública

	No se puede controlar quién recibe la intervención	No se puede evitar que el grupo de control pueda recibir el tratamiento	Todos los individuos reciben la intervención	Preocupación porque no se complete la evaluación de impacto	Recursos limitados para la realización de una evaluación de impacto	Dificultades en la determinación de las relaciones causales de la intervención	Responsables políticos son escépticos acerca de la intervención pública
Diseño experimental	No apropiada	No apropiada	No apropiada	Alto	Puede ser cara	Puede ser apropiada	Bastante apropiada
Regresión en discontinuidad	No apropiada	No apropiada	No apropiada	Alto	Puede ser cara	Puede ser apropiada	Bastante apropiada
Series temporales	Puede ser apropiada	No se basa en grupo de control	Puede ser apropiada	Bajo	Puede ser barata	Puede ser apropiada	Apropiada
Construcción de grupo de control para emparejamiento	Puede ser apropiada	Apropiada	No apropiada	Bajo	Puede ser barata	Puede ser apropiada	Bastante apropiada
Identificación causal y eliminación exhaustiva	Puede ser apropiada	No hay grupo de control formal	Puede ser apropiada	Bajo	Puede ser barata	No apropiada	Poco apropiada
Juicio de expertos	Puede ser apropiada	No hay grupo de control formal	Puede ser apropiada	Bajo	Puede ser barata	No apropiada	Poco apropiada
Juicio de agentes clave	Puede ser apropiada	No hay grupo de control formal	Puede ser apropiada	Bajo	Puede ser barata	No apropiada	Poco apropiada

6.2. Etapa 2. Selección de la técnica de evaluación óptima en el caso de construcción de grupos de control para emparejamiento

De los diseños de evaluación de impacto previamente presentados, el diseño consistente en la construcción de un grupo de control para realizar emparejamiento con el de tratamiento se considera a menudo como un diseño muy pragmático, ya que puede ser utilizado en las situaciones en las que el evaluador no tiene control sobre quién recibe una intervención (situación muy habitual en los programas sociales). Este hecho de no permitir al evaluador controlar la asignación del tratamiento descarta la posibilidad de considerar la situación como diseño experimental puro o un diseño de regresión en discontinuidad. Sin embargo, aunque estos diseños son ampliamente aceptados desde un punto de vista ético, unido a que también suele ser viable su realización (por ejemplo, desde un punto de vista de asignación de la intervención), existe un importante problema que se debe tener presente, y es el relativo al alto riesgo de que el grupo de comparación que se utiliza sea diferente en aspectos importantes del grupo de tratamiento. Dependiendo de cuáles sean las características (observadas o no) en cada situación que hacen que los individuos de control y tratamiento se diferencien en su comportamiento, existen determinadas técnicas que pueden utilizarse para tratar estas diferencias entre el grupo de control y el grupo de intervención, mediante el uso de una de las siguientes cuatro opciones:

- Técnica de diferencias en diferencias.
- Técnica de variables instrumentales.
- *Propensity score matching*.
- Método de emparejamiento.

En la segunda etapa para determinar el procedimiento óptimo de evaluación de impacto, la tabla 4 nos ofrece un método de selección para determinar cuál de las técnicas de construcción de un grupo de control emparejado es mejor, y se aplica solo en aquellas situaciones en las que pueda haber dudas acerca del grupo de control equivalente a un grupo de tratamiento. En esta segunda tabla, el paso inicial consiste en recorrer la columna en la parte izquierda de la tabla para identificar los elementos que se aplican a la intervención gubernamental que se desea evaluar. A continuación, se selecciona la técnica de evaluación que corresponda a este programa público. Las características que se exponen en la columna de la izquierda de la toma de decisiones en la tabla 4 son:

- Es posible observar las tendencias temporales del grupo de control y de tratamiento. En esta situación se puede rastrear las tendencias por separado, tanto en el grupo de control como en el grupo de intervención.
- Existe información de una (o más) característica(s) (variables) que afectan a que un individuo decida participar en el programa público, es decir, forme parte del grupo de tratamiento, y que no tenga relación con el resultado logrado en la variable de interés. Tales características se pueden utilizar para identificar un subconjunto de los posibles miembros del grupo de control que son más propensos a ser como los miembros del grupo de tratamiento.
- Es posible describir al grupo de tratamiento y a los posibles miembros del grupo de control de manera precisa lo que permite hacer predicciones sobre el resultado probable de una persona (u otra unidad) en ausencia de la intervención. El hecho de disponer de datos con alta calidad tanto del grupo de tratamiento como de los posibles

miembros del grupo de control puede ser utilizado para que, mediante técnicas matemáticas, se pueda predecir los resultados esperados para alguien que no recibió la intervención pública.

- Se puede construir un grupo de control mediante la localización de otros individuos (o unidades) que son emparejados de manera exacta con los miembros del grupo de tratamiento en ciertas características clave. Esta técnica se puede utilizar para asegurar que los miembros del grupo de tratamiento y del grupo de control son lo suficientemente parecidos.

Tabla 4. Selección de técnicas de evaluación para mejorar el diseño de la construcción de un grupo de control para emparejamiento

Qué es posible en esta situación	Técnica	Cómo realizarla
¿Se puede realizar un seguimiento tanto del grupo de control como de tratamiento a lo largo del tiempo? Para un mismo año, el grupo de control comienza en un valor de la variable de interés diferente al que presenta el grupo de tratamiento, pero es posible seguir las tendencias tanto del grupo de comparación como las del grupo de intervención a lo largo del tiempo.	Diferencias en diferencias	Seguimiento tanto del grupo de control como de tratamiento y calcular que la mejoría de la variable de resultado en el grupo de tratamiento difiere considerablemente de la experimentada por el grupo de control.
¿Se puede encontrar una característica (variable) que, además de no estar relacionada con la variable de interés, haga que las unidades (individuos) quieran pertenecer al grupo de tratamiento? Entonces se puede usar esta variable para crear un subconjunto en el grupo de control muy semejante a los individuos del grupo de tratamiento, ya que la única razón por la que no están en el grupo de tratamiento es debido a esta característica. Por ejemplo, ellos viven muy lejos del lugar donde se realiza la intervención. Por lo tanto, no pertenecen al grupo de tratamiento, no por cuestiones de motivación (que están relacionados con la var. impacto) sino que se debe solo al costo de transporte (incorrelada con la var. de interés).	Variable instrumental	Comparar los resultados del grupo de tratamiento con los de un subconjunto de individuos que potencialmente pertenecen al grupo de control —solo se tendrá en cuenta aquellos que tienen la característica de selección (por ejemplo, no van a la universidad porque viven a bastante distancia) y suponiendo que no están en el grupo de tratamiento debido (solo) a esta razón (es decir, vivir demasiado lejos)— y que son similares a los del grupo de tratamiento en el resto de las otras variables importantes.

Qué es posible en esta situación	Técnica	Cómo realizarla
¿Es posible describir tanto al grupo de tratamiento como a los posibles miembros del grupo de control de manera muy precisa, lo que permite realizar estimaciones sobre cuál puede ser el resultado más probable que presente un individuo, en el caso de no haber recibido la intervención? Este resultado estimado (en ausencia de la intervención) se compara entonces con el resultado real que se produjo para el grupo de tratamiento.	Propensity score matching	Seleccionar el grupo potencial de control (es decir, aquel que no han recibido ningún tipo de intervención) y mediante procedimientos econométricos y estadísticos, intentar estimar los resultados a partir de variables explicativas o características de dicho grupo (por ejemplo, edad, sexo, educación, raza, discapacidad). Calcular una ecuación que permita estimar el resultado de las unidades que presentan ciertas características particulares. Para cada uno de los miembros del grupo de tratamiento utilizar esta misma fórmula para estimar el valor de la variable de resultados que probablemente habría obtenido en el caso de no recibir la intervención. Comparar los resultados reales en la variable de resultado (es decir, la observada realmente después de haber recibido la intervención) con los resultados estimados (es decir, lo que habría ocurrido si no hubieran recibido la intervención).
¿Es capaz de formar un grupo de control mediante la localización de otros individuos (o unidades) que "coinciden" exactamente con los miembros del grupo de tratamiento en las características clave (variables)?	Emparejamiento	Para cada miembro del grupo de tratamiento, se localiza a individuos que tienen características similares a estos miembros del grupo de intervención, salvo por el hecho de no haber recibido la intervención. Finalmente, se comparan los resultados de los miembros del grupo de tratamiento con la variable de resultado observada en su "emparejado" en el grupo de control.

6.2.1. Guía de decisión de la técnica de evaluación de impacto utilizando grupos de control

6.2.1.1. Selección del método de evaluación

La gran preocupación a la hora de seleccionar el método de evaluación es la manera en la que se abordará el problema de sesgo de selección. Cómo se hará esto depende en gran medida de la comprensión que se tenga sobre el sesgo, que a su vez, requiere un buen entendimiento sobre el comportamiento de los beneficiarios del programa. En la figura 9 se muestra un árbol de decisión para la selección de un enfoque de evaluación. Los pasos básicos de este árbol de decisión son los siguientes:

1. Si la evaluación se diseña de manera ex-ante, la primera pregunta que se debe realizar es si es posible la asignación aleatoria. Si el grupo de tratamiento se selecciona de forma aleatoria, entonces, la extracción del grupo de tratamiento solucionaría todos los problemas existentes de sesgos, y se podría considerar un grupo de comparación válido. Esto no significa que este enfoque tenga que afectar a toda la población, ya que es posible focalizar este diseño, por ejemplo, en hogares pobres que no superan un determinado índice de pobreza.

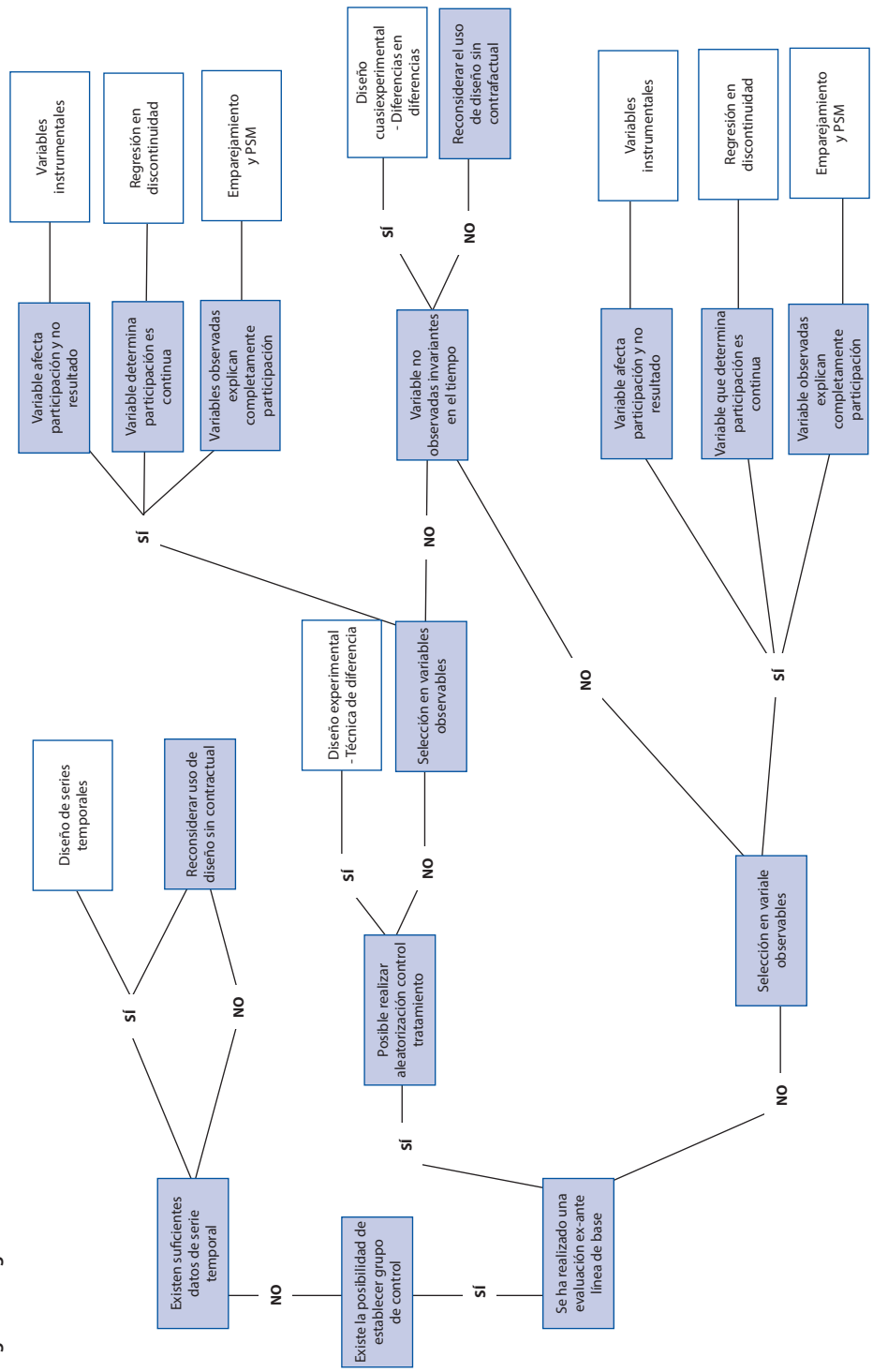


Figura 9. Regla de decisión del diseño de evaluación cuantitativa

2. Si hay sesgos, ¿son observables todos factores determinantes de la participación? Si es así, entonces hay una serie de enfoques basados en técnicas de regresión que puede eliminar el sesgo de selección.
3. Si los determinantes de selección son no observados, pero se cree que son invariantes en el tiempo, un diseño de "diferencias en diferencias" y utilizando datos de panel se podría eliminar su influencia, por lo que es fundamental disponer de datos antes de que se implemente la política, es decir, una línea de base.
4. Si el estudio es ex-post no se dispone de un panel, por lo que hay un problema de selección que está determinado por las características no observables, entonces debería buscarse algún medio que ofrezca alguna información sobre cómo son esos supuestos no observables. Si no es posible, será necesario emplear diseños con lista de espera si hay beneficiarios que aún no han sido tratados.
5. Si no es posible ninguna de las opciones anteriores, entones el problema de sesgo de selección no se puede abordar, y habría que considerar la imposibilidad de aplicar diseños con contrafactual, ya que cualquier evaluación de impacto tendrá que depender en gran medida de la teoría del programa y supuestos difícilmente contrastables.

7. Conclusiones

Este documento propone una guía metodológica que permita determinar la técnica de evaluación cuantitativa óptima, en términos de idoneidad, viabilidad, y que resulte asequible en términos presupuestarios, dependiendo de las características del programa implementado por el gobierno. Por lo tanto, los dos elementos fundamentales desarrollados en este manual son:

- En primer lugar, identificar criterios para definir si corresponde hacer una evaluación cuantitativa en una determinada intervención, estudiando qué condiciones se deben cumplir para que esa política sea merecedora de una evaluación de impacto.
- En segundo lugar, si se ha decido realizar una evaluación de impacto, se definen las etapas y criterios que guían el diseño de la evaluación de impacto más adecuado para la intervención a evaluar (previo a la elección de la técnica de análisis específica).

Bibliografía

Abadie, A., Angrist, J. D. e Imbens, G. W. (2002). "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings". *Econometrica,* 70 (1): 91-117.

Angrist, J. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administration Records". *American Economic Review*, 80 (3): 313-335.

Angrist, J., Bettinger, E., Bloom, E., King, E. y Kremer, M. (2002). "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment". *American Economic Review,* 92 (5): 1535-1558.

Bamberger, M. (1986). "Monitoring and evaluating urban development programs: a handbook for program managers and researchers". Washington DC: World Bank.

Banerjee, A., Cole, S., Duflo, E. y Linden, L. (2007). "Remedying Education: Evidence from Two Randomized Experiments in India". *Quarterly Journal of Economics,* 122 (3): 1235-1264.

Behrman, J. y Hoddinott, J. (2005). "Programme Evaluation with Unobserved Heterogeneity and Selective Implementation: The Mexican 'PROGRESA' Impact on Child Nutrition". *Oxford Bulletin of Economics and Statistics,* 67 (4): 547-569.

Bell, B., Blundell, R. y Reenen, van J. (1999). "Getting the Unemployed Back to Work: An Evaluation of the New Deal Proposals". *International Tax and Public Finance,* 6 (3): 339-360.

Bertrand, M., Dufl, E. y Mullainathan, S. (2004). "How Much Should We Trust Differences-in-Differences Estimates?". *Quarterly Journal of Economics,* 119 (1): 249-275.

Blundell, R. y Costa Días, M. (2008). "Alternative Approaches to Evaluation in Empirical Microeconomics". CeMMAP Working Paper 26/08. London: Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Bryk, A. S. (ed.). (1983). *Stakeholder-Based Evaluation*. San Francisco: Joseey-Bass.

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

Card, D. y Krueger, A. B. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *The American Economic Review*, 84 (4).

Caliendo, M. y Kopeinig, S. (2008). "Some Practical Guidance for the Implementation" of Chen, S., Mu, R. y Ravallion, M. (2008). "Are There Lasting Impacts of Aid to Poor Areas? Evidence for Rural China". *Policy Research Working Paper 4084.* Washington, DC: World Bank.

Callon, M., Laredo P. y Mustar, P. (1995). *La gestion stratgique de la recherche et de la technologie,* (31-88). Paris: Economica.

Chaudhury, N. y Parajuli, D. (2006). "Conditional Cash Transfers and Female Schooling: The Impact of the Female School Stipend Program on Public School Enrollments in Punjab, Pakistan". Policy Research Working Paper 4102. Washington DC: World Bank.

Chen, H. T. (1990). "Theory-Driven evaluations". Newbury Park, CA: SAGE.

Cozzens, S. E. (1987). "Expert Review in Evaluating Programs". *Science and Public Policy*, 14 (2): 71-81.

Dehejia, R. (2005). "Practical Propensity Score Matching: A Reply to Smith and Todd". *Journal of Econometrics,* 125 (1-2): 355-364.

Duflo, E., Glennerster, R. y Kremer, M. (2008). "Using Randomization in Development Economics Research: A Toolkit". In (ed. T. Paul Schultz y J. Strauss), *Handbook of Development Economics*, 4: 3895-3962. Amsterdam: North-Holland.

Duignan, P. (2009). "Impact/outcome evaluation designs and techniques illustrated with a simple example". *Outcomes Theory Knowledge Base*, nº 237.

Galasso, E. y Ravallion, M. (2004). "Social Protection in a Crisis: Argentina's Plan *Jefes y Jefas*", *World Bank Economic Review,* 18 (3): 367-400.

Gertler, P., Martínez, S., Premad, P., Rawlings, L. B. y Vermeersch, C. M. J. (2011). "Impact evaluation in practice". Washington DC: World Bank.

Gruber, J. (1994). "The Incidence of Mandated Maternity Benefits". *American Economic Review* 84 (3): 622-641.

Hahn, J., Todd, P. y Klaauw, van der W. (2001). "Identifi cation of Treatment Effects by Regression Discontinuity Design". *Econometrica,* 69 (1): 201-209.

Hahn, J., Hirano, K. y Karlan, D. (2008). "Adaptive Experimental Design Using the Propensity Score". Working Paper 969. New Haven, CT: Economic Growth Center, Yale University.

Harvey, A. C. (1990) (2ª ed.). "The Econometric Analysis of Time Series", vol. 1. The MIT Press.

Heckman, J. J. (1997). "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations". *Journal of Human Resources,* 32 (3): 441-462.

Heckman, J. J., Ichimura, H. y Todd, P. (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies,* 64 (4): 605-654.

Heckman, J. J. y Vytlacil, E. (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". *Econometrica,* 73 (3): 669-738.

Healely, P. (1998). "Collaborative Planning in a stakeholder society". *Town Planning Review*, 69 (1).

Holden, D. y Zimmerman, M. (2009). "A practical guide to program evaluation planning". London: SAGE publications.

Khandker, S., Koolwal, G. B. y Samad, H. A. (2010). "Handbook of impact evaluation: quantitative methods and practices". Washington DC: World Bank.

Kusek, J. Z. y Rist, R. C. (2004). *A Handbook for Development Practitioners: Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: World Bank.

Lope-Acevedo, G. y Tan, H. W. (2010). "Impact evaluation of SME programs in Latin America and Caribbean". Washington, DC: World Bank.

Nadeau, M. A. (1988). *L'évaluation de programme Laval*, (349-352). Québec: Presses de l'université de Laval.

National Science Fundation (2002). "The 2002 user friendly handbook for project evaluation". NSF.

ODA (1995). *Guidance Note on Indicators for Measuring and Assessing Primary Stakeholder Participation*. ODA Social Development Department, July 1995.

Ravillion, M. (2008). "Evaluating Anti-Poverty Programs". In (ed. T. Paul Schultz y J. Strauss), *Handbook of Development Economics*, 4: 3787-3846. Amsterdam: North-Holland.

Weiss, C. (1998). *Evaluation*. New Jersey: Prentice Hall.

Wholey, J., Hatry, H. y Newcomer, K. (2010). "Handbook of practical program evaluation". San Francisco: Wiley.

Witkin, B. R. y Altschuld, J. W. (1995). *Planning Conducting Needs Assessments*, (193-203). Thousand Oaks: Sage Publications.

Anexo I. Secuencia en la realización de una evaluación

A la hora de realizar una evaluación de un programa público, las fases que hay que realizar son las siguientes:

1. Preparación de la evaluación.
   a) Decidir qué se quiere evaluar.
   b) Objetivos, cuestiones de política pública.
   c) Desarrollo de hipótesis, marcos lógicos, cadenas de resultados.
   d) Selección de indicadores.

2. Planificar y desarrollar el diseño de evaluación.
   a) Elección del diseño de evaluación.
   b) Confirmar que esa evaluación es ética.
   c) Crear y preparar un grupo de evaluación.
   d) Momento temporal en que se realizará la evaluación.
   e) Presupuesto de la evaluación.

3. Selección de la muestra.
   a) Decidir el tamaño de la muestra.
   b) Decidir la estrategia de muestreo.

4. Recopilar los datos.
   a) Decidir qué tipo de datos se necesitan recopilar.
   b) Contratar ayuda para la recopilación de información.
   c) Desarrollar un cuestionario.
   d) Encuesta piloto.
   e) Realizar trabajo de campo.
   f) Procesamiento de la información y validación de datos.

5. Producción y publicación de resultados.
   a) Analizar los datos.
   b) Escribir el informe.
   c) Debatir los resultados con los gestores políticos.
   d) Publicitar los resultados.

Anexo II. 'Checklist' de las evaluaciones

Lista de validación de cada una de las técnicas de evaluación

A continuación se presenta una tabla para cada una de las técnicas de evaluación analizadas en este documento que analiza la idoneidad, viabilidad y asequibilidad de cada una de ellas.

'Checklist' del diseño experimental

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce de un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada.
- ✓ ¿Es ético no dar la intervención al grupo de control?
- ✓ ¿Es posible medir la variable de resultado del programa con la suficiente precisión?

'Checklist' de viabilidad

- ✓ ¿Existe un número lo suficiente grande de potenciales participantes (unidades)? La población debe ser lo suficientemente grande como para disponer de un número considerable de participantes (unidades) en el momento de hacer la evaluación.
- ✓ Bajas tasas de deserción. Existe una proporción elevada de individuos, tanto del grupo de control como de tratamiento, que continúa durante todo el experimento.
- ✓ ¿El tamaño final de la muestra es lo suficientemente grande? El tamaño final de la muestra tiene que ser lo suficientemente grande como para que los análisis estadísticos sean robustos.
- ✓ ¿El grupo de control no recibe ninguna intervención compensatoria? El grupo de control no puede recibir ningún tipo de intervención compensatoria durante el transcurso del experimento, ya que existe la posibilidad de que tanto ellos, como grupos de interés externos, puedan querer compensaciones por la imposibilidad de recibir la intervención original.
- ✓ ¿Se ha suministrado la intervención pública de forma adecuada? Necesidad de garantizar que la intervención se ha realizado (entregado) de forma correcta. Para que se pueda considerar una estimación creíble del efecto de la política es necesario que se tenga el suficiente control tanto sobre la calidad como sobre la cantidad de la intervención que se realiza.
- ✓ ¿Se realiza la recopilación de información (datos) de forma correcta? Necesidad de tener suficiente control sobre la metodología y forma en que se recogen los datos para que estos sean una medición apropiada de la variable de impacto que se desea estudiar.
- ✓ ¿Existe una descripción apropiada del "proceso" de la intervención? Necesidad de describir el proceso y el contexto del programa de modo que sea posible interpretar resultados de

las evaluaciones negativas (por ejemplo, puede ser que los problemas que existen en la variable de resultados sean debidos a que la intervención se ha "entregado" de forma inadecuada).

'Checklist' de asequibilidad

- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, el control de procesos de intervención, recolección de datos, análisis de datos, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta su conclusión.

'Checklist' del diseño de regresión en discontinuidad

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce de un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada.
- ✓ ¿Es ético no dar la intervención al grupo de control? La técnica de regresión en discontinuidad resulta más ética porque el programa se implementa a aquellos individuos que más lo necesitan.
- ✓ ¿Es posible medir la variable que determina el tratamiento para un número considerable de valores (que sea continua)? Se necesita que la variable de selección que establece quién recibe y no el tratamiento tenga suficientes valores como para ser considerada continua.

'Checklist' de viabilidad

- ✓ ¿Existe un número lo suficiente grande de potenciales participantes (unidades)? La población debe ser lo suficientemente grande como para disponer de un número considerable de participantes (unidades) en el momento de hacer la evaluación.
- ✓ Bajas tasas de deserción. Existe una proporción elevada de individuos, tanto de grupo de control como de tratamiento, que continúa durante todo el experimento.
- ✓ ¿El tamaño final de la muestra es lo suficientemente grande? El tamaño final de la muestra tiene que ser lo suficientemente grande para que los análisis estadísticos sean robustos.
- ✓ ¿El grupo de control no recibe ninguna intervención compensatoria? El grupo de control no puede recibir ningún tipo de intervención compensatoria durante el transcurso del experimento, ya que existe la posibilidad de que tanto ellos, como grupos de interés externos, puedan querer compensaciones por la imposibilidad de recibir la intervención original.
- ✓ ¿Se ha suministrado la intervención pública de forma adecuada? Necesidad de garantizar que la intervención se ha realizado (entregada) de forma adecuada. Para que se

pueda considerar una estimación correcta del efecto de la política es necesario que se tenga el suficiente control tanto sobre la calidad como sobre la cantidad de la intervención que se realiza.

- ✓ ¿Se realiza la recopilación de información (datos) de forma correcta? Necesidad de tener suficiente control sobre la metodología y forma en que se recogen los datos para que estos sean una medición apropiada de la variable de impacto que se desea estudiar.
- ✓ ¿Existe una descripción apropiada del "proceso" de la intervención? Necesidad de describir el proceso y el contexto del programa de modo que sea posible interpretar resultados de las evaluaciones negativas (por ejemplo, puede ser que los problemas que existen en la variable de resultados sean debidos a que la intervención se ha "entregado" de forma inadecuada)

'Checklist' de asequibilidad

- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, el control de procesos de intervención, recolección de datos, análisis de datos, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta llegar a su conclusión.

'Checklist' del diseño de series temporales

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce de un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada**.**
- ✓ ¿Es un programa con carácter universal? Es decir, ¿se aplica la intervención a todo el mundo? El análisis de series temporales se puede utilizar en el caso de que la política pública se aplique a toda una población.

'Checklist' de viabilidad

- ✓ ¿Se dispone de un número de observaciones lo suficientemente amplio? ¿Va a ser posible tener un número lo suficientemente grande de mediciones de la variable de interés?
- ✓ Medición precisa del momento en que se realiza la intervención pública. ¿Es posible medir de forma precisa el momento de tiempo en que se introdujo la intervención de manera que se sea posible establecer una relación entre este momento temporal con los valores observados en las series temporales de la variable de resultados?
- ✓ ¿Se ha suministrado la intervención pública de forma adecuada? Necesidad de garantizar que la intervención se ha realizado (entregada) de forma adecuada. Para que se pueda considerar una estimación correcta del efecto de la política es necesario que se tenga

el suficiente control tanto sobre la calidad como sobre la cantidad de la intervención que se realiza.

- ✓ ¿Se realiza la recopilación de información (datos) de forma correcta? Necesidad de tener suficiente control sobre la metodología y forma en que se recogen los datos para que estos sean una medición apropiada de la variable de impacto que se desea estudiar.
- ✓ ¿Existe una descripción apropiada del "proceso" de la intervención? Necesidad de describir el proceso y el contexto del programa de modo que sea posible interpretar resultados de las evaluaciones negativas (por ejemplo, puede ser que los problemas que existen en la variable de resultados sean debidos a que la intervención se ha "entregado" de forma inadecuada).

'Checklist' de asequibilidad

- ✓ ¿La medición es continua y suficientemente larga a lo largo del tiempo? ¿Se dispone de suficientes fuentes de información para tener series temporales lo suficientemente largas?
- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, el control de procesos de intervención, recolección de datos, análisis de datos, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta su conclusión.

'Checklist' del diseño de grupos de control para emparejamiento

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce de un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada.
- ✓ ¿Es ético no dar la intervención al grupo de control?
- ✓ ¿El grupo de control es lo suficientemente similar al grupo de tratamiento? Es necesario disponer de suficiente información de los individuos del grupo de control de tal forma que se garantice que se pueden emparejar completamente con los del grupo de tratamiento. En el caso de existir variables distintas en los individuos que impliquen que quieran participar en el programa y que no puedan ser medidas en el grupo de control (sesgo en variable no observada) esto puede ser un problema.
- ✓ ¿Existen las mismas condiciones que operan en diferentes periodos de tiempo? En el caso de que la información relativa al grupo de control venga de un periodo de tiempo distinto no debería existir nada significativo en el momento temporal en que se toman los datos del grupo de control respecto al momento en que se tomó la información del grupo de tratamiento.

'Checklist' de viabilidad

- ✓ ¿Se dispone de un número de individuos del grupo de control lo suficientemente amplio?
- ✓ ¿No hay intervención de compensación en el grupo de control? El grupo de control no puede recibir ningún tipo de intervención compensatoria a lo largo del experimento de evaluación, ya que tanto ellos, como otros agentes, pueden desear compensarlos por el hecho de no recibir el tratamiento.
- ✓ Bajas tasas de abandono. Proporción de individuos suficientemente grande en el grupo de tratamiento y de control que continúa durante toda la evaluación.
- ✓ ¿La muestra final es lo suficientemente grande? Se necesita que el tamaño muestral sea lo suficientemente grande para que tenga robustez en términos estadísticos.
- ✓ ¿Se ha suministrado la intervención pública de forma adecuada? Necesidad de garantizar que la intervención se ha realizado (entregada) de forma adecuada. Para que se pueda considerar una estimación correcta del efecto de la política es necesario que se tenga el suficiente control tanto sobre la calidad como sobre la cantidad de la intervención que se realiza.
- ✓ ¿Se realiza la recopilación de información (datos) de forma correcta? Necesidad de tener suficiente control sobre la metodología y forma en que se recogen los datos para que estos sean una medición apropiada de la variable de impacto que se desea estudiar.
- ✓ ¿Existe una descripción apropiada del "proceso" de la intervención? Necesidad de describir el proceso y el contexto del programa de modo que sea posible interpretar los resultados de las evaluaciones negativas.

'Checklist' de asequibilidad

- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, el control de procesos de intervención, recolección de datos, análisis de datos, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta su conclusión.

'Checklist' del diseño de causal y eliminación exhaustiva

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada**.**
- ✓ ¿Existe una definición clara de "éxito" del programa? Es necesario que esté claro a partir de qué valor de la variable de interés se considera que el programa ha tenido éxito.
- ✓ ¿Existe información lo suficientemente rica sobre cómo se implementa el programa y los mecanismos de causalidad? Es deseable que sea posible recopilar información rica sobre los posibles mecanismo a través de los cuales otros posibles factores pueden afectar a la variable de interés (necesario que el evaluador sea capaz de disponer de suficiente

información que le permita hacer juicios justificados sobre mecanismos de causalidad alternativos).

'Checklist' de viabilidad

- ✓ ¿Se dispone de un número de participantes en el estudio? ¿Existen suficientes participantes en el estudio en el que la variable de interés fue positiva a lo largo del periodo de evaluación?
- ✓ ¿Se realiza la recopilación de información (datos) de forma correcta? Necesidad de tener suficiente control sobre la metodología y forma en que se recogen los datos para que estos sean una medición apropiada de la variable de impacto que se desea estudiar.
- ✓ ¿Es posible medir de forma precisa la variable de resultado? ¿Se ha medido de forma adecuada la variable de impacto en aquellos que reciben el tratamiento? En el caso de que la medición sea correcta, es necesario poder examinar qué causas han generado el valor de la variable de interés, para así poder eliminar otras explicaciones de causalidad alternativas. Los casos en los que no se han alcanzado los valores esperados también deben ser analizados.
- ✓ ¿Es posible identificar causas alternativas? Es necesario que exista la posibilidad de identificar exhaustivamente todas las posibles causas alternativas.
- ✓ ¿Se eliminan las causas alternativas? Es necesario disponer de suficiente información que permita eliminar las causas alternativas que pueden afectar al logro de los objetivos en la variable de interés.
- ✓ ¿Existe una descripción apropiada del "proceso" de la intervención? Necesidad de describir el proceso y el contexto del programa de modo que sea posible interpretar resultados de las evaluaciones negativas.

'Checklist' de asequibilidad

- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, el control de procesos de intervención, recolección de datos, análisis de datos, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta su conclusión.

'Checklist' del diseño de juicio de expertos

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada**.**
- ✓ ¿Se pregunta —y contesta— la cuestión relevante en la evaluación? Si se contesta a una pregunta incorrecta es un desperdicio de tiempo y esfuerzo.

- ✓ ¿Existen expertos reconocidos en esa área a los que se puede solicitar que realicen una evaluación?
- ✓ ¿Los expertos pueden realizar juicios del programa? Es un tema en el que los expertos son capaces de hacer un juicio sobre si el programa ha mejorado los resultados. Esto requiere por un lado que los mecanismos causales estén lo suficientemente claros para los expertos y que, además, no existan demasiadas alternativas que puedan haber afectado a los resultados.
- ✓ ¿Los actores clave van a aceptar esto como diseño suficientemente robusto? Las principales partes interesadas tienen que aceptar que este método de evaluación les proporcionará resultados suficientemente robustos comparando el impacto del programa para los fines para los que quieran utilizar estos resultados y el nivel de recursos que tienen que gastar en la evaluación.

'Checklist' de viabilidad

- ✓ ¿Los expertos están suficientemente libres de prejuicios? ¿Será posible localizar expertos que sean suficientemente libres de sesgo (ya sea a favor o en contra del programa) de manera que puedan proporcionar un juicio fiable sobre los resultados del programa?
- ✓ ¿Las opiniones de expertos serán aceptadas por las partes interesadas? ¿Se las considera un juicio fiable sobre los resultados del programa y serán aceptados por un número suficiente de actores clave?

'Checklist' de asequibilidad

¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, la selección de los participantes, contratar expertos, pagar la recopilación de la información que los expertos necesitarán para realizar un juicio, solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta su conclusión.

'Checklist' del diseño de juicio de informantes clave

'Checklist' de idoneidad

- ✓ ¿Se ha realizado —y contestado— la pregunta relevante de la evaluación? Sobre la base de lo que ya se conoce un determinado sector y los temas prioritarios de evaluación necesarios para hacerse una idea lo más completa y correcta de un sector.
- ✓ ¿Es el mejor proyecto sobre el que responder a la pregunta de evaluación? Es posible que exista un proyecto mejor en ese sector (o más habitual, o más innovador, o que resulte más fácil de controlar) en el que responder a la pregunta de evaluación realizada, de tal forma que se compartan y combinen recursos de evaluación de diferentes proyectos para responder de forma adecuada**.**
- ✓ ¿Se pregunta —y contesta— la cuestión relevante en la evaluación? Si se contesta a una pregunta incorrecta es un desperdicio de tiempo y esfuerzo.
- ✓ ¿Los informantes clave son capaces de realizar juicios del programa? Es un tema en el que los expertos son capaces de hacer un juicio sobre si el programa ha mejorado los resultados. Esto requiere por un lado que los mecanismos causales estén lo suficientemente claros

para los expertos y que, además, no existan demasiadas alternativas que puedan haber afectado a los resultados.

- ✓ ¿Los actores clave van a aceptar esto como diseño suficientemente robusto? Las principales partes interesadas tienen que aceptar que este método de evaluación les proporcionará resultados suficientemente robustos comparando el impacto del programa para los fines para los que quieran utilizar estos resultados y el nivel de recursos que tienen que gastar en la evaluación.

'Checklist' de viabilidad

- ✓ ¿Existen informantes clave dispuestos a participar en la toma de decisiones? ¿Habrá una gama suficientemente amplia de informantes clave dispuestos a hacer juicios sobre si el programa mejora los resultados?
- ✓ ¿Los informantes clave están suficientemente libres de prejuicios? ¿Será posible localizar expertos que sean suficientemente libre de sesgo (ya sea a favor o en contra del programa) de manera que puedan proporcionar un juicio fiable sobre los resultados del programa?
- ✓ ¿Los juicios de los informantes clave serán aceptados por las partes interesadas? ¿Serán considerados juicios fiables sobre los resultados del programa y serán aceptados por un número suficiente de actores clave?

'Checklist' de asequibilidad

- ✓ ¿Es posible financiar hasta el final toda la evaluación? Es necesario que existan suficientes recursos (para la gestión y administración de la evaluación, los grupos de interés, entrevistar a los informantes claves solución de problemas, informes y difusión de resultados) para continuar con la evaluación hasta llegar a su conclusión.

Anexo III. Ejemplo de aplicación de este procedimiento

Ejemplo de adecuación, viabilidad y asequibilidad análisis de la evaluación de impacto

En este anexo se va a realizar una ilustración que muestra un análisis sobre cada una de las diferentes técnicas de evaluación de impacto o de resultados de largo plazo identificadas en este documento, para estudiar si cada una de ellas se puede utilizar en función de un análisis de la pertinencia, la viabilidad y la asequibilidad de la evaluación en un programa público. Se examina cada uno de los diseños de evaluación de impacto identificados en las secciones anteriores, mostrando conclusiones en términos de su adecuación, viabilidad y asequibilidad. Como ya se ha mencionado, hay que tener en cuenta que los dos últimos diseños de evaluación de impacto identificados en esta teoría de resultados (opinión de los expertos y los diseños de juicios y opiniones de informantes clave) a menudo son rechazados por algunos grupos, ya que no presentan el suficiente apoyo empírico y cuantitativo para determinar de manera clara si existe una relación de causalidad entre programa y resultado. Sin embargo, se ha decido incluirlos debido a que existen algunos sectores interesados que consideran que pueden ser adecuados para algunas situaciones.

Ejemplo. Creación de una nueva ley a nivel nacional sobre requisitos en la construcción de vivienda nueva

Se pretende realizar un plan de evaluación para saber el efecto que ha tenido la implementación de una nueva ley sobre criterios técnicos en la construcción de nueva edificación. Se ha decido introducir un nueva ley nacional debido al fracaso (se produjeron grietas y fugas de agua) en bastantes edificios que se habían construido bajo el anterior régimen legal. A continuación se presenta el análisis de los posibles diseños de evaluación de impacto:

Diseño experimental

No es factible. Este diseño consiste en establecer una comparación entre un grupo que recibe la intervención y un grupo que no la recibe (idealmente seleccionados al azar). Parece claro que, por motivos éticos, políticos y jurídicos no es posible aplicar esta nueva ley (llevar a cabo las intervenciones) solo en una o más localidades, utilizando a otras zonas (que sirven como grupo de control) ya que no se ven afectadas por la nueva reglamentación. Entre otras cosas, la nueva regulación no puede ser impuesta solo una parte del país. Además, existe un grave problema del diseño en la práctica, ya que es más que probable que se produzca una reacción en aquellas regiones en las que no se aplica la nueva ley (sacando otra ley parecida), estableciéndose de este modo una rivalidad compensatoria que reduciría el efecto final en la variable de resultados entre el grupo de tratamiento y el grupo de control.

Diseño de regresión discontinuidad

No es factible. Este diseño asume que es posible ordenar (y representar) a las localidades de un país que podrían recibir la intervención en función de una variable continua medible (por ejemplo, la calidad de los edificios en la localidad). Entonces, la intervención solo se aplicaría a aquellas localidades por debajo de un cierto nivel de corte. Cualquier efecto debe mostrar un desplazamiento hacia arriba de la gráfica alrededor del punto de corte. En teoría, es posible clasificar a las

ciudades en función de la calidad de su obra nueva y suponiendo que los recursos disponibles para realizar la intervención fueran limitados entonces resultaría ético intervenir solo en aquellas ciudades con la peor obra nueva, y, por lo tanto, establecer una discontinuidad en el diseño de la regresión. Sin embargo, el compromiso legal así como el diseño de la propia política (como en el diseño experimental anterior) significa que un diseño de regresión en discontinuidad no es factible.

Diseño de series temporales

No es factible. En este diseño se mide un resultado un gran número de veces a lo largo del tiempo, y luego se examina si se produce un cambio claro en el momento de tiempo donde se introdujo la intervención. Este diseño sería posible si se dispone de mediciones consecutivas sobre la calidad de las edificaciones nuevas a lo largo de una serie temporal larga (digamos 20 años). Sin embargo, este diseño tiene el problema de que se produzca un *shock* en el mismo momento de tiempo en que entra en vigor la ley —supongamos que este nuevo factor es un "compromiso de responsabilidad" firmado por todos los constructores del país—. Este compromiso surge como consecuencia de que todos los constructores se dan cuenta de que pueden ser denunciados por los compradores de pisos debido a fallos en los edificios. Cabe señalar que este "compromiso de responsabilidad" no significa que no se pueda analizar la serie temporal como un modo de seguimiento del indicador. El único problema es que al realizar este tipo de análisis de series temporales no somos capaces de estimar si la razón por la que se produce el cambio es debida a la nueva ley o a ese "compromiso de responsabilidad" presentado por los constructores.

Diseño de construcción de grupo de control para emparejamiento

No es factible. Este diseño consiste en establecer una comparación entre un grupo que recibe la intervención y un grupo que no la recibe (idealmente seleccionados al azar). Dado que tiene un carácter universal, no es posible localizar ningún grupo de individuos que pueda ser considerado un buen contrafactual del grupo que recibe el tratamiento.

Diseño de identificación causal y eliminación exhaustiva

Viabilidad baja. Este diseño funciona del siguiente modo: primero se identifica que se haya producido un cambio en la variable de resultados observados y, posteriormente, se lleva a cabo un análisis detallado de todas las posibles causas (alternativas a la aprobación de la nueva ley) que han generado el cambio en el resultado, para lograr su posterior eliminación. En algunos casos, se puede elaborar una lista detallada de las posibles causas que afectan a los resultados observados y luego usar un proceso para identificar cuál de ellos es más probable que haya causado el efecto observado. Sin duda, este enfoque va mucho más allá de la acumulación de evidencia para tratar de explicar el resultado observado debido solo a la intervención y requiere que todas las posibles explicaciones alternativas que puedan causar el resultado sean eliminadas. Esto puede no ser posible en este caso debido a la aparición simultánea del "documento de responsabilidad", que ya se comentó anteriormente, que se produjo en el mismo marco temporal que la intervención. Es posible que esta causa se entremezcle con la intervención original, por lo que al calcular el impacto de la medida no se sabe qué porcentaje es debido a la nueva ley y cuál al "documento de responsabilidad".

Diseño de opinión de expertos

Viabilidad alta. Este diseño consiste en pedir opinión a un experto en la materia para analizar una situación y evaluar si, ponderando todos los elementos de los que disponen, aceptan la hipótesis de que la nueva ley pueda haber causado el resultado. Es necesario preguntar a varios expertos independientes en la regulación de la construcción de edificio, incluso expertos extranjeros con el fin de garantizar la independencia), y que puedan visitar el país y de este modo evaluar si los cambios en la calidad de los nuevos edificios son debidos al nuevo régimen regulatorio. Estos juicios se basan en su criterio profesional y tendrían en cuenta todos aquellos datos que necesitan para hacer sus afirmaciones. En sus informes se explicaría por qué y de qué modo se llegó a esta decisión. Este enfoque es muy factible, pero proporciona un nivel significativamente más bajo de certeza que todos los otros diseños de evaluación de impacto descritos anteriormente. En el caso de usar este diseño, la pregunta de evaluación que se responde siempre debe estar claramente identificada, como por ejemplo: ¿en la opinión de un experto independiente la nueva ley de construcción ha producido una mejora en los resultados de la construcción nueva? También resulta necesario realizar un estudio de factibilidad para analizar en detalle las posibilidades de este tipo de diseño.

Diseño de opinión de informantes clave

Viabilidad alta. Este diseño consiste en pedir a ciertos informantes clave (personas que tienen acceso por razón de su cargo al conocimiento de lo que ha ocurrido con respecto a la intervención) que analicen si creen que la intervención puede haber causado el resultado buscado. Es necesario realizar una selección de aquellas partes interesadas más importantes y que deseen actuar como informantes, y que mediante la realización de entrevistas reflejen sus opiniones con respecto a los resultados que pueden ser atribuidos al nuevo régimen jurídico.

EUROsociAL es un programa de cooperación regional de la Unión Europea con América Latina para la promoción de la cohesión social, mediante el apoyo a políticas públicas nacionales, y el fortalecimiento de las instituciones que las llevan a cabo. EUROsociAL pretende promover un diálogo euro-latinoamericano de políticas públicas en torno a la cohesión social. Su objetivo es contribuir a procesos de reforma e implementación en diez áreas clave de políticas, en ciertas temáticas, seleccionadas por su potencial impacto sobre la cohesión social. El instrumento del que se dota es el de la cooperación institucional o aprendizaje entre pares: el intercambio de experiencias y la asesoría técnica entre instituciones públicas de Europa y de América Latina.

Consorcio Liderado por

FIIAPP
COOPERACIÓN ESPAÑOLA

PROGRAMA FINANCIADO POR LA UNIÓN EUROPEA

Socios Coordinadores

APC Colombia · ENAP · FEI · giz · iila

Participan más de 80 Socios Operativos y Entidades Colaboradoras de Europa y América Latina

Este documento propone una guía metodológica que permita determinar la técnica óptima de evaluación de impacto de una política pública concreta, en términos de idoneidad y viabilidad, y que resulte asequible desde el punto de vista presupuestario. Inicialmente se presentan las posibles actividades de evaluación y seguimiento que se pueden realizar en una intervención pública, centrando el estudio en la evaluación cuantitativa y de impacto. Una vez descritas las técnicas más habituales existentes en la evaluación de impacto, los dos elementos fundamentales desarrollados en este manual son, en primer lugar, la identificación de criterios que permitan definir si corresponde hacer una evaluación cuantitativa en una determinada intervención del Estado, estudiando qué condiciones se deben cumplir para que esa política sea merecedora de una evaluación de impacto. En segundo lugar, en el caso de haber decido realizar una evaluación de impacto, se definen las etapas y criterios que guían el diseño de la evaluación de impacto más adecuado para la intervención a evaluar (previo a la elección de la técnica de análisis específica).

www.eurosocial-ii.eu