La evaluación de impacto en la práctica

SEGUNDA EDICIÓN

Paul J. Gertler, Sebastián Martínez,
Patrick Premand, Laura B. Rawlings
y Christel M. J. Vermeersch

GRUPO BANCO MUNDIAL

BID
Banco Interamericano de Desarrollo

La evaluación de impacto en la práctica

Segunda edición

Se recomienda visitar el sitio web del libro *La evaluación de impacto en la práctica* en http://www.worldbank.org/ieinpractice. El sitio web contiene materiales de apoyo, e incluye soluciones para las preguntas del estudio de caso HISP del libro, así como la correspondiente base de datos y el código de análisis del *software* Stata; un manual técnico que proporciona un tratamiento más formal del análisis de datos; presentaciones de PowerPoint relacionadas con los capítulos; una versión en línea del libro con hipervínculos a los sitios web, y enlaces con otros materiales.

Este libro ha sido posible gracias al generoso apoyo del Fondo Estratégico de Evaluación de Impacto (SIEF, por sus siglas en inglés: *Strategic Impact Evaluation Fund*). Lanzado en 2012 con apoyo del Departamento para el Desarrollo Internacional del Reino Unido, el SIEF es un programa de alianzas que promueve la elaboración de políticas públicas basadas en la evidencia. Actualmente se centra en cuatro ámbitos cruciales para un desarrollo humano saludable: educación básica, sistemas de salud y prestación de servicios, desarrollo infantil temprano y nutrición, y agua y saneamiento. El SIEF funciona en todo el mundo, sobre todo en países de bajos ingresos, aportando conocimientos expertos sobre la evaluación de impacto, y evidencia para diversos programas y equipos de elaboración de políticas públicas.

La evaluación de impacto en la práctica

Segunda edición

Paul J. Gertler, Sebastián Martínez,
Patrick Premand, Laura B. Rawlings
y Christel M. J. Vermeersch

GRUPO BANCO MUNDIAL

BID
Banco Interamericano de Desarrollo

© 2017 Banco Internacional para la Reconstrucción y el Desarrollo/Banco Mundial
1818 H Street NW, Washington, DC 20433
Teléfono: 202-473-1000; Internet: www.worldbank.org
Algunos derechos reservados

1 2 3 4 20 19 18 17

Los hallazgos, interpretaciones y conclusiones recogidas en esta obra no reflejan necesariamente el punto de vista del Banco Mundial ni de su Directorio Ejecutivo, del Banco Interamericano de Desarrollo, de su Directorio Ejecutivo ni de los gobiernos que representan. El Banco Mundial y el Banco Interamericano de Desarrollo no garantizan la precisión de los datos incluidos en esta obra. Las fronteras, colores, denominaciones y otra información de cualquier mapa en esta obra no implican ningún juicio por parte del Banco Mundial ni el Banco Interamericano de Desarrollo en relación con el estatus legal de cualquier territorio ni la aprobación ni aceptación de dichas fronteras.

Ningún contenido de esta obra constituirá o será considerado como una limitación ni renuncia de los privilegios e inmunidades del Banco Mundial ni del Banco Interamericano de Desarrollo, privilegios e inmunidades específicamente reservados.

Derechos y permisos

Esta obra está disponible bajo la licencia de atribución de Creative Commons 3.0 IGO (CC BY 3.0 IGO) http://creativecommons.org/licenses/by/3.0/igo. En el marco de la licencia de atribución Creative Commons, se permite copiar, distribuir, transmitir y adaptar esta obra, incluso para objetivos comerciales, bajo las siguientes condiciones:

Atribución: se ruega citar la obra de la siguiente manera: Gertler, Paul J., Sebastián Martínez, Patrick Premand, Laura B. Rawlings y Christel M. J. Vermeersch. 2017. *La evaluación de impacto en la práctica*, Segunda edición. Washington, DC: Banco Interamericano de Desarrollo y Banco Mundial. doi:10.1596/978-1-4648-0888-3. Licencia de atribución: Creative Commons CC BY 3.0 IGO

Traducciones: Si se procede a una traducción de esta obra, se ruega añadir la siguiente exención de responsabilidad con la atribución: *Esta traducción no es una creación del Banco Mundial y no se debería considerar una traducción oficial del Banco Mundial. El Banco Mundial no será responsable de ningún contenido o error en esta traducción.*

Adaptaciones: Si se crea una adaptación de esta obra, se ruega añadir la siguiente exención de responsabilidad o con la siguiente atribución: *Ésta es una adaptación de una obra original del Banco Mundial. Las ideas y opiniones expresadas en la adaptación son responsabilidad exclusiva del autor o de los autores de la adaptación y no han sido refrendadas por el Banco Mundial.*

Contenidos de terceros: El Banco Mundial no es dueño necesariamente de cada componente del contenido de este trabajo. Por lo tanto, el Banco Mundial no garantiza que el uso de cualquier componente individual o parte propiedad de terceros contenido en la obra no vulnerará los derechos de esos terceros. El riesgo de reclamaciones que resulten de dicha vulneración incumbe solo a usted. Si quiere reutilizar un componente de la obra, es responsabilidad suya determinar si se requiere una autorización para esa reutilización y para obtener permiso del dueño de los derechos de autor. Los ejemplos de los componentes pueden incluir cuadros, gráficos o imágenes, si bien no están limitados a ellos.

Todas las consultas sobre derechos y licencias deberán dirigirse a la División de Publicación y Conocimiento, Banco Mundial, 1818 H Street NW, Washington, DC 20433, EE.UU.; fax: 202-522-2625; e-mail: pubrights@worldbank.org.

ISBN (papel): 978-1-4648-0888-3
ISBN (electrónica): 978-1-4648-0989-7
DOI: 10.1596/978-1-4648-0888-3

Ilustración: C. Andrés Gómez-Peña y Michaela Wieser
Diseño de la tapa: Critical Stages

CONTENIDOS

Prólogo xv

Agradecimientos xxi

Acerca de los autores xxiii

Abreviaturas xxvii

PRIMERA PARTE. INTRODUCCIÓN A LA EVALUACIÓN DE IMPACTO 1

Capítulo 1. ¿Por qué evaluar? 3
La formulación de políticas basada en evidencia 3
¿Qué es la evaluación de impacto? 7
Evaluación de impacto prospectiva versus evaluación retrospectiva 10
Estudios de eficacia y estudios de efectividad 12
Enfoques complementarios 14
Consideraciones éticas con respecto a la evaluación de impacto 22
La evaluación de impacto en las decisiones de políticas 24
La decisión de llevar a cabo una evaluación de impacto 29

Capítulo 2. La preparación de una evaluación 35
Pasos iniciales 35
Construcción de una teoría del cambio 36
Desarrollo de una cadena de resultados 38
La especificación de las preguntas de la evaluación 40
La selección de indicadores de resultados y desempeño 46
Lista de verificación: datos para los indicadores 47

SEGUNDA PARTE. CÓMO EVALUAR 51

Capítulo 3. Inferencia causal y contrafactuales 53
Inferencia causal 53

El contrafactual 55
Dos estimaciones falsas del contrafactual 60

Capítulo 4. La asignación aleatoria 71
La evaluación de programas basados en reglas de asignación 71
La asignación aleatoria del tratamiento 72
Lista de verificación: la asignación aleatoria 91

Capítulo 5. Las variables instrumentales 99
La evaluación de programas cuando no todos cumplen su asignación 99
Tipos de estimaciones de impacto 100
El cumplimiento imperfecto 102
Promoción aleatoria como variable instrumental 112
Lista de verificación: promoción aleatoria como variable instrumental 122

Capítulo 6. Diseño de regresión discontinua 125
Evaluación de programas que utilizan un índice de elegibilidad 125
El diseño de regresión discontinua difuso 131
Verificación de la validez del diseño de regresión discontinua 132
Limitaciones e interpretaciones del método de diseño de regresión discontinua 137
Lista de verificación: diseño de regresión discontinua 139

Capítulo 7. Diferencias en diferencias 143
Evaluación de un programa cuando la regla de asignación es menos clara 143
El método de diferencias en diferencias 144
¿Qué utilidad tiene el método de diferencias en diferencias? 148
El supuesto de "tendencias iguales" en el método de diferencias en diferencias 150
Limitaciones del método de diferencias en diferencias 156
Verificación: diferencias en diferencias 156

Capítulo 8. Pareamiento 159
Construcción de un grupo de comparación artificial 159
Pareamiento por puntajes de propensión 161
La combinación del pareamiento con otros métodos 164
Limitaciones del método de pareamiento 173
Verificación: el pareamiento 174

Capítulo 9. Cómo abordar las dificultades metodológicas 177
Efectos heterogéneos del tratamiento 177
Efectos no intencionados en la conducta 178
Imperfección del cumplimiento 179
El efecto de derrame 181
El desgaste 188
Programación en el tiempo y persistencia de los efectos 191

Capítulo 10. Evaluación de programas multifacéticos 195
Evaluación de programas que combinan diversas opciones de tratamiento 195
Evaluación de programas con diferentes niveles de tratamiento 196
Evaluación de múltiples intervenciones 199

TERCERA PARTE. CÓMO IMPLEMENTAR UNA EVALUACIÓN DE IMPACTO 205

Capítulo 11. Elección de un método de evaluación de impacto 207
¿Qué método usar en un determinado programa? 207
Cómo las reglas operativas de un programa pueden contribuir a elegir un método de evaluación de impacto 208
Una comparación de métodos de evaluación de impacto 214
Cómo encontrar la unidad de intervención más pequeña factible 218

Capítulo 12. Gestión de una evaluación de impacto 223
Gestión del equipo, del tiempo y del presupuesto de una evaluación 223
Roles y responsabilidades de los equipos de investigación y de políticas públicas 224
Establecer una colaboración 231
Cómo programar una evaluación en el tiempo 237
Cómo elaborar un presupuesto para una evaluación 240

Capítulo 13. La ética y la ciencia en la evaluación de impacto 257
La gestión de evaluaciones éticas y creíbles 257
La ética de llevar a cabo evaluaciones de impacto 258
Garantizar evaluaciones fiables y creíbles mediante la ciencia abierta 264
Lista de verificación: una evaluación de impacto ética y creíble 271

Capítulo 14. Divulgación de resultados y generación de impacto en las políticas públicas 275
Una base de evidencia sólida para las políticas públicas 275

Elaboración a la medida de una estrategia de comunicación para diferentes públicos 279
Divulgación de los resultados 283

CUARTA PARTE. CÓMO OBTENER DATOS PARA UNA EVALUACIÓN DE IMPACTO 289

Capítulo 15. La elección de una muestra 291
El muestreo y los cálculos de potencia 291
Elaboración de una muestra 291
La decisión sobre el tamaño de la muestra de una evaluación de impacto: cálculos de potencia 297

Capítulo 16. Encontrando fuentes adecuadas de datos 323
Tipos de datos necesarios 323
La utilización de datos cuantitativos existentes 326
La recopilación de datos de nuevas encuestas 332

Capítulo 17. Conclusiones 355
Las evaluaciones de impacto: ejercicios complejos pero valiosos 355
Lista de verificación: elementos centrales de una evaluación de impacto bien diseñada 356
Lista de verificación: recomendaciones para mitigar riesgos habituales al llevar adelante una evaluación de impacto 357

Glosario 361

Recuadros

1.1 Cómo una evaluación exitosa puede promover la sostenibilidad política de un programa de desarrollo 5
1.2 El impacto de las políticas de un modelo preescolar innovador 7
1.3 Pruebas de la capacidad generalizable de los resultados 13
1.4 Simulación de posibles efectos del proyecto a través del modelado estructural 16
1.5 Un método mixto de evaluación en acción 17
1.6 Fundamentos para una ampliación a escala nacional mediante una evaluación de procesos en Tanzania 19
1.7 La evaluación de costo-efectividad 21
1.8 Evaluación de programas innovadores 25
1.9 La evaluación de alternativas de diseño de programas 26
1.10 El enfoque de evaluaciones de impacto de *clusters* 28

2.1 La articulación de una teoría del cambio: de los pisos de cemento a la felicidad en México 37
2.2 Experimentos de mecanismo 41
2.3 Una reforma de las matemáticas en la enseñanza secundaria: elaboración de una cadena de resultados y una pregunta de la evaluación 43
3.1 El problema del contrafactual: la "señorita Única" y el programa de transferencias condicionadas 56
4.1 La asignación aleatoria como un valioso instrumento operativo 73
4.2 La asignación aleatoria como regla de selección de un programa: las transferencias condicionadas y la educación en México 78
4.3 Asignación aleatoria de donaciones para mejorar las perspectivas de empleo juvenil en el norte de Uganda 79
4.4 Asignación aleatoria de intervenciones en abastecimiento de agua y saneamiento en zonas rurales de Bolivia 79
4.5 Asignación aleatoria de protección del agua de pozos para mejorar la salud en Kenia 80
4.6 Asignación aleatoria e información a propósito de los riesgos del VIH para reducir el embarazo adolescente en Kenia 81
5.1 El uso de variables instrumentales para evaluar el impacto de *Plaza Sésamo* en la preparación escolar 101
5.2 Variables instrumentales para lidiar con la falta de cumplimiento en un programa de vales escolares en Colombia 110
5.3 Promoción de inversiones en infraestructura educativa en Bolivia 118
6.1 Uso del diseño de regresión discontinua para evaluar el impacto de la reducción de las tarifas escolares en los índices de matriculación en Colombia 126
6.2 Redes de protección social basadas en un índice de pobreza en Jamaica 130
6.3 El efecto en el desempeño escolar de la agrupación de alumnos según sus puntuaciones en las pruebas educativas en Kenia 133
7.1 Utilización del método DD para entender el impacto de los incentivos electorales en las tasas de abandono escolar en Brasil 145
7.2 Aplicación del método de diferencias en diferencias para estudiar los efectos del despliegue policial en la tasa de delitos en Argentina 149
7.3 Comprobando el supuesto de tendencias iguales: privatización del agua y mortalidad infantil en Argentina 153

7.4 Poniendo a prueba el supuesto de tendencias iguales: la construcción de escuelas en Indonesia 154
8.1 Diferencias en diferencias pareadas: caminos rurales y desarrollo del mercado local en Vietnam 165
8.2 Pareamiento de diferencias en diferencias: suelos de cemento, salud infantil y felicidad de las madres en México 166
8.3 El método de control sintético: los efectos económicos de un conflicto terrorista en España 168
9.1 Cuentos tradicionales de la evaluación de impacto: el efecto Hawthorne y el efecto John Henry 178
9.2 Externalidades negativas debidas a efectos de equilibrio general: asistencia para la colocación laboral y resultados del mercado de trabajo en Francia 183
9.3 Trabajando con los efectos de derrame: remedios antiparasitarios, externalidades y educación en Kenia 184
9.4 Evaluación de los efectos de derrame: transferencias condicionadas y derrames en México 187
9.5 El desgaste en estudios con seguimiento a largo plazo: desarrollo infantil temprano y migración en Jamaica 189
9.6 Evaluación de los efectos a largo plazo: subsidios y adopción de redes antimosquitos tratadas con insecticidas en Kenia 191
10.1 Prueba de la intensidad de un programa para mejorar la adhesión a un tratamiento antirretroviral 198
10.2 Pruebas de alternativas de los programas para monitorear la corrupción en Indonesia 199
11.1 Programas de transferencias monetarias condicionadas y el nivel mínimo de intervención 221
12.1 Principios rectores de la participación de los equipos de políticas públicas y de evaluación 228
12.2 Descripción general de un plan de evaluación de impacto 229
12.3 Ejemplos de modelos de equipos de investigación y de políticas públicas 234
13.1 Registro de pruebas en las ciencias sociales 267
14.1 El impacto en las políticas públicas de un modelo innovador de educación preescolar en Mozambique 277
14.2 Instrumentos de extensión y divulgación 284
14.3 La divulgación efectiva de las evaluaciones de impacto 285
14.4 Divulgación de las evaluaciones de impacto en línea 286
14.5 Blogs de evaluación de impacto 287
15.1 El muestreo aleatorio no es suficiente para la evaluación de impacto 295

16.1 Elaboración de una base de datos en la evaluación del Plan Nacer de Argentina 330
16.2 Utilización de datos censales para reevaluar el PRAF en Honduras 331
16.3 Diseño y formato de los cuestionarios 338
16.4 Algunas ventajas y desventajas de la recopilación electrónica de datos 342
16.5 Recopilación de datos para la evaluación de las pruebas piloto de atención a crisis en Nicaragua 348
16.6 Directrices para la documentación y el almacenamiento de datos 349

Gráficos

2.1 Los elementos de una cadena de resultados 39
B2.2.1 Identificación de un experimento de mecanismo en una cadena de resultados más larga 42
B2.3.1 Cadena de resultados para la reforma de la currícula de matemática en la escuela secundaria 43
2.2 La cadena de resultados del HISP 45
3.1 El clon perfecto 57
3.2 Un grupo de comparación válido 59
3.3 Estimaciones antes-después de un programa de microfinanzas 61
4.1 Características de los grupos bajo tratamiento con asignación aleatoria 76
4.2 Muestra aleatoria y asignación aleatoria de tratamiento 81
4.3 Pasos para la asignación aleatoria del tratamiento 85
4.4 Asignación aleatoria del tratamiento mediante hoja de cálculo 87
4.5 Estimación del impacto con la asignación aleatoria 90
5.1 Asignación aleatoria con cumplimiento imperfecto 106
5.2 Estimación del efecto local promedio del tratamiento bajo asignación aleatoria con cumplimiento imperfecto 107
5.3 Proceso de promoción aleatoria 116
5.4 Estimación del efecto local promedio del tratamiento bajo la promoción aleatoria 117
6.1 Producción de arroz, fincas pequeñas vs. fincas grandes (línea de base) 128
6.2 Producción de arroz, fincas pequeñas vs. fincas grandes (seguimiento) 129
6.3 Cumplimiento de la asignación 132
6.4 Manipulación del índice de elegibilidad 133

6.5 HISP: densidad de los hogares, según el índice de pobreza de línea de base 135
6.6 Participación en el HISP, según el índice de pobreza de línea de base 135
6.7 Índice de pobreza y gastos en salud: el HISP dos años después 136
7.1 El método de diferencias en diferencias 146
7.2 Diferencias en diferencias cuando las tendencias de los resultados son diferentes 151
8.1 Pareamiento exacto en cuatro características 160
8.2 Pareamiento por puntajes de propensión y rango común 162
8.3 Pareamiento para el HISP: rango común 170
9.1 Un ejemplo clásico de efecto de derrame: externalidades positivas de la administración de remedios antiparasitarios a los niños de las escuelas 186
10.1 Pasos para la asignación aleatoria de dos niveles de tratamiento 197
10.2 Pasos para la asignación aleatoria de dos intervenciones 200
10.3 Diseño híbrido para un programa con dos intervenciones 201
15.1 Uso de una muestra para inferir las características promedio de una población de interés 292
15.2 Un marco muestral válido cubre el conjunto de la población de interés 293
B15.1.1 Muestreo aleatorio entre grupos no comparables de participantes y no participantes 296
B15.1.2 Asignación aleatoria de los beneficios de un programa entre un grupo de tratamiento y un grupo de comparación 297
15.3 Una muestra más grande tiene más probabilidades de parecerse a la población de interés 300

Cuadros

3.1 Evaluación del HISP según comparación antes-después 64
3.2 Impacto del HISP según comparación antes-después (análisis de regresión) 64
3.3 Evaluación del HISP según comparación inscritos-no inscritos (comparación de medias) 67
3.4 Evaluación del HISP según comparación inscritos-no inscritos (análisis de regresión) 68
4.1 Evaluación del HISP: balance entre los pueblos de tratamiento y de comparación en la línea de base 93
4.2 Evaluación del HISP según la asignación aleatoria (comparación de medias) 94

4.3	Evaluación del HISP según la asignación aleatoria (análisis de regresión)	95
5.1	Evaluación del HISP según la promoción aleatoria (comparación de medias)	120
5.2	Evaluación del HISP según la promoción aleatoria (análisis de regresión)	121
6.1	Evaluación del HISP: diseño de regresión discontinua con análisis de regresión	137
7.1	Cálculo del método de diferencias en diferencias	147
7.2	Evaluación del HISP: diferencias en diferencias (comparación de medias)	155
7.3	Evaluación del HISP: diferencias en diferencias (análisis de regresión)	155
8.1	Estimación del puntaje de propensión a partir de características observables de la línea de base	169
8.2	Evaluación del HISP: pareamiento de las características de línea de base y comparación de medias	171
8.3	Evaluación del HISP: pareamiento de las características de línea de base y análisis de regresión	171
8.4	Evaluación del HISP: método de diferencias en diferencias combinado con pareamiento en las características de línea de base	172
B10.1.1	Resumen del diseño del programa	198
11.1	Relación entre las reglas operativas de un programa y los métodos de evaluación de impacto	211
11.2	Comparación de métodos de evaluación de impacto	215
12.1	Costo de las evaluaciones de impacto de una selección de proyectos con apoyo del Banco Mundial	241
12.2	Costos desagregados de una selección de proyectos con apoyo del Banco Mundial	242
12.3	Ejemplo de presupuesto para una evaluación de impacto	250
13.1	Asegurar información fiable y creíble para las políticas mediante la ciencia abierta	265
14.1	Participación de grupos clave en el impacto en las políticas: por qué, cuándo y cómo	280
15.1	Ejemplos de *clusters*	304
15.2	Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables, potencia = 0,9	309

15.3 Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables, potencia = 0,8 310
15.4 Evaluación del HISP+: tamaño requerido de la muestra para detectar diversos efectos mínimos deseados (aumento de la tasa de hospitalización) 311
15.5 Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables (disminución de los gastos del hogar en salud) 314
15.6 Evaluación del HISP+: tamaño requerido de la muestra para detectar un impacto mínimo de US$2 en diversas cantidades de *clusters* 315

PRÓLOGO

Este libro ofrece una introducción accesible al tema de la evaluación de impacto y su práctica en el desarrollo. Proporciona orientaciones provechosas para diseñar e implementar evaluaciones de impacto, junto con una visión general no técnica de los métodos de evaluación de impacto.

Esta es la segunda edición del manual de *La evaluación de impacto en la práctica*. Publicado por primera vez en 2011, el manual ha sido usado extensamente por comunidades de desarrollo y académicas en todo el mundo. La primera edición está disponible en inglés, francés, portugués y español.

La versión actualizada abarca las técnicas más recientes para evaluar programas e incluye consejos de implementación de última generación, así como un conjunto ampliado de ejemplos y estudios de casos que se basan en recientes intervenciones de desarrollo. También incluye nuevos materiales sobre la ética de la investigación y alianzas para llevar a cabo evaluaciones de impacto. A lo largo del libro, los estudios de casos ilustran aplicaciones de las evaluaciones de impacto. La publicación incluye enlaces de material didáctico complementario disponible en línea.

El enfoque de la evaluación de impacto que se vuelca en este libro es sobre todo intuitivo. Hemos intentado minimizar las anotaciones técnicas. Los métodos se basan directamente en la investigación aplicada en ciencias sociales y comparten numerosos elementos comunes a los métodos de investigación que se emplean en las ciencias naturales. En este sentido, la evaluación de impacto reúne herramientas de investigación empírica ampliamente utilizadas en economía y en otras ciencias sociales, junto con las realidades operativas y de economía política de la implementación de políticas públicas y práctica de desarrollo.

Nuestro enfoque de la evaluación de impacto también es pragmático: pensamos que deberían definirse los métodos más apropiados para adecuarse al contexto operativo, y no a la inversa. La mejor manera de lograr

esto es al comienzo de un programa, mediante el diseño de evaluaciones de impacto prospectivas que se incluyan en la implementación de un proyecto. Sostenemos que alcanzar un consenso entre las principales partes interesadas y la definición de un diseño de evaluación adecuado al contexto político y operativo es tan importante como el propio método. También creemos que las evaluaciones de impacto deberían ser claras a propósito de sus limitaciones y sus advertencias. Por último, alentamos encarecidamente a los responsables de las políticas públicas y a los administradores de los programas a considerar las evaluaciones de impacto como parte de una teoría bien desarrollada del cambio que establece con claridad las vías causales mediante las cuales un programa funciona para elaborar productos e influir en los resultados finales, y los alentamos a combinar las evaluaciones de impacto con enfoques de monitoreo y de evaluación complementarios con el fin de obtener un cuadro completo de los resultados.

Nuestras experiencias y lecciones sobre cómo llevar a cabo evaluaciones de impacto en la práctica se basan en la enseñanza y en el trabajo con cientos de socios idóneos de los ámbitos gubernamentales, académicos y del desarrollo. El libro se fundamenta colectivamente en décadas de experiencia en el trabajo con evaluaciones de impacto en casi todos los rincones del planeta, y está dedicado a las futuras generaciones de profesionales y responsables de las políticas públicas.

Esperamos que estas páginas constituyan un valioso recurso para la comunidad internacional de desarrollo, las universidades y los responsables de las políticas públicas que intentan construir evidencia válida en torno a lo que funciona en el desarrollo. Más y mejores evaluaciones de impacto contribuirán a fortalecer la base de evidencia para las políticas y los programas de desarrollo en todo el mundo. Tenemos la esperanza de que si los gobiernos y los profesionales del desarrollo pueden tomar decisiones de políticas públicas sobre la base de la evidencia, incluida la evidencia generada a través de la evaluación de impacto, los recursos para el desarrollo se destinarán de manera más efectiva para reducir la pobreza y mejorar las vidas de las personas.

Hoja de ruta de los contenidos del libro

En la primera parte, “Introducción a la evaluación de impacto” (capítulos 1 y 2), se explica por qué una evaluación de impacto puede llevarse a cabo y cuándo merece la pena hacerlo. Allí se revisan los diversos objetivos que una evaluación de impacto puede lograr y se subrayan las principales preguntas relativas a las políticas públicas que una evaluación puede abordar. Se insiste

en la necesidad de definir rigurosamente una teoría del cambio que explique los canales a través de los cuales los programas pueden influir en los resultados finales. Se insta a una consideración cuidadosa de los indicadores de resultados y del tamaño de los efectos anticipados.

En la segunda parte, "Cómo evaluar" (capítulos 3 al 10), se analizan las diversas metodologías que producen grupos de comparación que se pueden utilizar para estimar los impactos de un programa. Se empieza por introducir el *contrafactual* como la piedra angular de cualquier evaluación de impacto, explicando las propiedades que debe tener la estimación del mismo, y proporcionando ejemplos de estimaciones inválidas del contrafactual. Posteriormente, se presenta un menú de las opciones de evaluación de impacto que pueden producir estimaciones válidas del contrafactual. En particular, se aborda la intuición básica que subyace a las metodologías de evaluación de impacto, a saber: la *asignación aleatoria*, las *variables instrumentales*, el *diseño de regresión discontinua*, las *diferencias en diferencias* y el *pareamiento*. Se establece por qué y cómo cada método puede producir una estimación válida del contrafactual, en qué contexto de las políticas públicas se puede implementar cada uno, así como también sus principales limitaciones.

A lo largo de esta parte del libro, se utiliza un estudio de caso –el Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés: *Health Insurance Subsidy Program*)– para ilustrar cómo se pueden aplicar los métodos. Además, se ofrecen ejemplos específicos de las evaluaciones de impacto que han utilizado cada método. La segunda parte concluye con un debate sobre cómo combinar los métodos y abordar problemas que pueden surgir durante la implementación, reconociendo que los diseños de evaluación de impacto a menudo no se implementan exactamente como se había planeado originalmente. En este contexto, se analizan ciertos problemas comunes que suelen experimentarse durante la implementación, lo que incluye el cumplimiento imperfecto o los efectos de derrame, y se debate cómo abordar estas dificultades. El capítulo 10 concluye con orientaciones sobre evaluaciones de programas multifacéticos, sobre todo aquellos con diferentes niveles de tratamiento y diseños cruzados.

La tercera parte, "Cómo implementar una evaluación de impacto" (capítulos 11 a 14), se centra precisamente en cómo llevar adelante la evaluación. En el capítulo 11, se detalla cómo utilizar las reglas del funcionamiento de un programa –es decir, los recursos disponibles del programa, los criterios para seleccionar a los beneficiarios y la programación en el tiempo para la implementación– como la base para seleccionar un método de evaluación de impacto. Se define un marco sencillo para determinar cuál de las metodologías de evaluación de impacto presentadas en la

segunda parte es la más adecuada para un determinado programa, de acuerdo con sus reglas operativas. En el capítulo 12 se aborda la relación entre el equipo de investigación y el equipo de políticas públicas, y sus respectivos roles para conformar conjuntamente un equipo de evaluación. Se examina la diferencia entre independencia y ausencia de sesgo, y se ponen de relieve ámbitos que pueden ser delicados para llevar a cabo una evaluación de impacto. Se ofrece orientación sobre cómo gestionar las expectativas, se destacan algunos de los riesgos habitualmente presentes en la realización de evaluaciones de impacto, y se brindan sugerencias sobre cómo manejarlos. El capítulo concluye con una visión general de cómo gestionar las actividades de la evaluación de impacto, lo que incluye la creación de un equipo de evaluación, la programación en el tiempo de la misma, el presupuesto, la captación de fondos y la recopilación de datos. En el capítulo 13 se proporciona una visión general de la ética y la ciencia de la evaluación de impacto, lo cual incluye la importancia de no negar beneficios a los beneficiarios elegibles en aras de la evaluación; en el capítulo también se resalta el rol de las juntas de revisión institucional, que aprueban y monitorean la investigación con sujetos humanos, y se aborda la importancia de registrar las evaluaciones siguiendo la práctica de la ciencia abierta, de acuerdo con la cual los datos se ponen a disposición del público para posteriores investigaciones y para replicar resultados. El capítulo 14 proporciona una visión novedosa sobre cómo utilizar las evaluaciones de impacto para fundamentar las políticas públicas, incluyendo consejos sobre cómo conseguir que los resultados sean relevantes; un debate sobre el tipo de productos que las evaluaciones de impacto pueden y deben producir, y orientación sobre cómo extraer y divulgar las conclusiones para maximizar el impacto de las políticas públicas.

La cuarta parte, "Cómo obtener datos para una evaluación de impacto" (capítulos 15 a 17), se ocupa de la forma de recopilar datos, lo que incluye elegir la muestra y determinar el tamaño apropiado de la muestra de la evaluación (capítulo 15), así como también encontrar fuentes de datos adecuados (capítulo 16). El capítulo 17 concluye y proporciona algunas listas de verificación.

Material complementario en línea

En el sitio web de la evaluación de impacto en la práctica se ofrecen materiales de apoyo (http://www.worldbank.org/ieinpractice), incluyendo soluciones a las preguntas de los estudios de casos del HISP, la correspondiente

base de datos y el código de análisis del *software* Stata, así como un manual técnico que proporciona un tratamiento más formal del análisis de datos. Los materiales también abarcan presentaciones de PowerPoint relacionadas con los capítulos, y versiones en línea del libro con hipervínculos a sitios web y enlaces con otros materiales.

El sitio web de evaluación de impacto en la práctica también brinda vínculos con materiales relacionados con el Fondo Estratégico de Evaluación de Impacto (SIEF, por sus siglas en inglés) del Banco Mundial, la iniciativa Desarrollo de la Evaluación de Impacto (DIME, por sus siglas en inglés), de la misma institución, y sitios web de manuales de evaluación de impacto, así como el portal de evaluación de impacto del Banco Interamericano de Desarrollo (BID) y el curso de métodos de evaluación de impacto aplicados en la Universidad de California, Berkeley.

Desarrollo de *La evaluación de impacto en la práctica*

La primera edición del libro *La evaluación de impacto en la práctica* se basaba en un conjunto central de materiales didácticos desarrollados para los seminarios sobre "Cómo convertir las promesas en evidencia", organizados por la Oficina del Economista Jefe para Desarrollo Humano, en asociación con unidades regionales y con el Grupo de Investigación en Economía del Desarrollo del Banco Mundial. En el momento de redactar la primera edición, el seminario se había celebrado más de 20 veces en todas las regiones del mundo.

Los seminarios, y tanto la primera como la segunda edición de este manual, han sido posibles gracias a las generosas ayudas del gobierno de España, del Departamento para el Desarrollo Internacional del Reino Unido (DFID) y de la Fundación del Fondo de Inversión para la Infancia (CIFF, Reino Unido) a través de contribuciones al SIEF. La segunda edición también se ha beneficiado del apoyo de la Oficina de Planificación Estratégica y Efectividad en el Desarrollo del BID.

Esta segunda edición ha sido puesta al día para abarcar las técnicas más actualizadas, así como consejos de implementación de última generación, siguiendo los progresos alcanzados en este campo en los últimos años. También hemos ampliado el conjunto de ejemplos y estudios de casos para reflejar aplicaciones de amplio espectro de la evaluación de impacto en las operaciones de desarrollo y destacar sus vínculos con las políticas públicas. Por último, hemos incluido aplicaciones de técnicas de evaluación de impacto con Stata, utilizando la base de datos del estudio de caso del HISP como parte del material complementario en línea.

AGRADECIMIENTOS

Los materiales didácticos sobre los que se basa este libro han experimentado numerosas versiones, y han sido enseñados por diversos y talentosos profesores, todos los cuales han dejado su impronta en los métodos y en el enfoque de la evaluación de impacto enunciados en el libro. Queremos agradecer y reconocer las contribuciones y los aportes sustanciales de diversos profesores que han participado en los seminarios en los que se basaba la primera edición, y que incluyen a Paloma Acevedo Alameda, Felipe Barrera, Sergio Bautista-Arredondo, Stefano Bertozzi, Barbara Bruns, Pedro Carneiro, Jishnu Das, Damien de Walque, David Evans, Claudio Ferraz, Deon Filmer, Jed Friedman, Emanuela Galasso, Sebastián Galiani, Arianna Legovini, Phillippe Leite, Gonzalo Hernández Licona, Mattias Lundberg, Karen Macours, Juan Muñoz, Plamen Nikolov, Berk Özler, Nancy Qian, Gloria M. Rubio, Norbert Schady, Julieta Trias, y Sigrid Vivo Guzmán. Agradecemos los comentarios realizados por nuestros revisores pares en la primera edición del libro (Barbara Bruns, Arianna Legovini, Dan Levy y Emmanuel Skoufias) y la segunda edición (David Evans, Francisco Gallego, Dan Levy y Damien de Walque), así como también las observaciones de Gillette Hall. Deseamos asimismo expresar nuestro agradecimiento por los esfuerzos de un talentoso equipo organizador, que incluye a Holly Balgrave, Theresa Adobea Bampoe, Febe Mackey, Silvia Paruzzolo, Tatyana Ringland, Adam Ross y Jennifer Sturdy.

Extendemos igualmente nuestro reconocimiento a todos los que participaron en las transcripciones del borrador del seminario de julio de 2009 realizado en Beijing, China, en el que se basan partes de este libro, especialmente a Paloma Acevedo Alameda, Carlos Asenjo Ruiz, Sebastian Bauhoff, Bradley Chen, Changcheng Song, Jane Zhang y Shufang Zhang. Reconocemos a Garret Christensen y a la Berkeley Initiative for Transparency in the Social Sciences, así como a Jennifer Sturdy y Elisa Rothenbühler por sus aportes al capítulo 13. También agradecemos a Marina

Tolchinsky y Kristine Cronin por su excelente apoyo en la investigación; a Cameron Breslin y Restituto Cárdenas por el respaldo en la programación; a Marco Guzmán y Martin Ruegenberg por el diseño de las ilustraciones, y a Nancy Morrison, Cindy A. Fisher, Fiona Mackintosh y Stuart K. Tucker por el apoyo editorial durante la producción de la primera y la segunda edición del libro.

Reconocemos y agradecemos el apoyo permanente y el entusiasmo por este proyecto de nuestros directivos en el Banco Mundial y el Banco Interamericano de Desarrollo, y especialmente al equipo del SIEF, entre ellos Daphna Berman, Holly Blagrave, Restituto Cárdenas, Joost de Laat, Ariel Fiszbein, Alaka Holla, Aliza Marcus, Diana-Iuliana Pirjol, Rachel Rosenfeld y Julieta Trias. Estamos sumamente agradecidos por el apoyo recibido de la administración del SIEF, incluyendo a Luis Benveniste, Joost de Laat y Julieta Trias. Agradecemos igualmente a Andrés Gómez-Peña y Michaela Wieser del Banco Interamericano de Desarrollo, y a Mary Fisk, Patricia Katayama, y Mayya Revzina, del Banco Mundial, por su ayuda con las comunicaciones y el proceso de publicación.

La producción de la segunda edición de este libro en su versión en español fue realizada por la Oficina de Planificación Estratégica y Efectividad en el Desarrollo del Banco Interamericano de Desarrollo. Reconocemos particularmente a Carola Álvarez y Arturo Galindo por su apoyo en esta iniciativa. Quedamos endeudados con Andrés Gómez-Peña y Michaela Wieser por su esfuerzo y dedicación en la coordinación del proceso de producción editorial de este volumen. De igual manera, agradecemos especialmente a Alberto Magnet por la traducción del libro al español, así como a Claudia M. Pasquetti, a cargo de la edición y lectura de pruebas en dicho idioma. Cabe también nuestro reconocimiento del aporte de los revisores técnicos de cada uno de los capítulos en español: Paloma Acevedo, Jorge Marcelo Franco Quincot, Gastón Gertner y Bibiana Taboada.

Por último, quisiéramos brindar nuestro reconocimiento a los participantes de los numerosos talleres, sobre todo los celebrados en Abiyán, Accra, Adís Abeba, Amán, Ankara, Beijing, Berkeley, Buenos Aires, Cairo, Ciudad de Panamá, Ciudad del Cabo, Cuernavaca, Dakar, Daca, Fortaleza, Katmandú, Kigali, Lima, Madrid, Managua, Manila, Ciudad de México, Nueva Delhi, Paipa, Pretoria, Rio de Janeiro, San Salvador, Santiago, Sarajevo, Seúl, Sofía, Túnez y Washington, D.C.

Gracias a su interés, a sus inteligentes preguntas y a su entusiasmo, hemos sido capaces de aprender paso a paso qué buscan los responsables de las políticas públicas en las evaluaciones de impacto. Esperamos que este libro refleje sus ideas.

ACERCA DE LOS AUTORES

Paul J. Gertler es profesor de economía en la cátedra Li Ka Shing de la Universidad de California, Berkeley, donde imparte clases en la Escuela de Negocios Haas y en la Escuela de Salud Pública. También es director científico del Centro para una Acción Global Efectiva en la Universidad de California. Se desempeñó como economista jefe en la Red de Desarrollo Humano del Banco Mundial entre 2004 y 2007 y en la Cátedra Fundadora de la Junta de Directores de la Iniciativa Internacional para la Evaluación de Impacto (3ie) entre 2009 y 2012. En el Banco Mundial, dirigió los trabajos para institucionalizar y desarrollar la escala de la evaluación de impacto para aprender acerca de aquello que funciona en el desarrollo humano. Ha sido investigador principal en diversas evaluaciones de impacto multisitio, entre ellas el programa de TCE, de México, Progresa-Oportunidades, y en el sistema de salud Remuneración por Desempeño, de Ruanda. Posee un doctorado en economía de la Universidad de Wisconsin, y se ha desempeñado como docente en Harvard, en RAND y en la State University of New York en Stony Brook.

Sebastián Martínez es economista principal en la Oficina de Planificación Estratégica y Efectividad en el Desarrollo del Banco Interamericano de Desarrollo (BID). Su trabajo se centra en el fortalecimiento de la base de evidencia y en la efectividad en el desarrollo de los sectores social y de infraestructura, incluyendo salud, protección social, mercados laborales, agua y saneamiento, y vivienda y desarrollo urbano. Dirige un equipo de economistas que lleva a cabo investigación sobre los impactos de los programas y políticas públicas del desarrollo, apoya la implementación de evaluaciones de impacto de las operaciones y trabaja en la mejora de capacidades para los clientes y el personal. Antes de integrarse al BID, trabajó seis años en el Banco Mundial, dirigiendo evaluaciones de programas sociales en América Latina y en África Subsahariana. Posee un

doctorado en economía de la Universidad de California, Berkeley, con una especialización en desarrollo y microeconomía aplicada.

Patrick Premand es economista senior en protección social y práctica global laboral en el Banco Mundial. Dirige el trabajo analítico y operativo sobre protección social y redes de protección; mercados laborales, empleo juvenil y capacidad emprendedora, así como también desarrollo infantil temprano. Su investigación se centra en construir evidencia sobre la efectividad de las políticas de desarrollo a través de evaluaciones de impacto de programas sociales y de desarrollo humano a gran escala. Ha ocupado diversos cargos en el Banco Mundial, lo cual incluye la Unidad de Economía del Desarrollo Humano de la región de África, la Oficina del Economista Jefe para Desarrollo Humano y la Unidad de Pobreza de la región de América Latina y el Caribe. Posee un doctorado en economía de la Universidad de Oxford.

Laura B. Rawlings es especialista líder en protección social en el Banco Mundial, y ostenta más de 20 años de experiencia en el diseño, la implementación y la evaluación de programas de desarrollo humano. Administra tanto las operaciones como la investigación, centrándose en el desarrollo de enfoques innovadores para sistemas de protección social efectivos y escalables en entornos de bajos recursos. Ha sido líder de equipo responsable de desarrollar la Estrategia de Protección Social y Laboral 2012-2022 del Banco Mundial y, anteriormente, administradora del Fondo Estratégico de Evaluación de Impacto (SIEF). También se desempeñó como líder del sector para desarrollo humano en Centroamérica, donde fue la responsable de gestionar las carteras de salud, educación y protección social del Banco Mundial. Comenzó su carrera en esta última institución, en el Grupo de Investigación sobre el Desarrollo, donde trabajó en los programas de evaluación de impacto de programas sociales. Ha trabajado en América Latina y el Caribe así como en África Subsahariana, dirigiendo numerosas iniciativas de proyectos de investigación en los ámbitos de transferencias condicionadas, empleo público, fondos sociales, desarrollo infantil temprano y sistemas de protección social. Antes de integrarse al Banco Mundial, trabajó en el Consejo para el Desarrollo de Ultramar, donde dirigió un programa educativo sobre temas de desarrollo para el personal en el Congreso de Estados Unidos. Ha publicado numerosos libros y artículos en el campo de la evaluación del desarrollo humano y es profesora adjunta en el Programa de Desarrollo Humano Global de la Universidad de Georgetown, Washington D.C.

Christel M. J. Vermeersch es economista senior en Práctica Global en Salud, Nutrición y Población en el Banco Mundial. Se ocupa de temas

relacionados con el financiamiento del sector salud, el financiamiento basado en resultados, monitoreo y evaluación, y evaluación de impacto. Anteriormente se desempeñó en los ámbitos de educación, desarrollo infantil temprano y capacidades. Ha sido coautora de estudios de evaluación de impacto para programas de financiamiento basados en resultados en Argentina y Ruanda, un seguimiento de largo plazo de un estudio de estimulación de la temprana infancia en Jamaica, así como del manual de evaluación de impacto en salud del Banco Mundial. Antes de sumarse a esta última institución, fue becaria de investigación posdoctoral en la Universidad de Oxford. Posee un doctorado en economía de la Universidad de Harvard.

ABREVIATURAS

3IE	Iniciativa Internacional para la Evaluación de Impacto
ATE	Efecto promedio del tratamiento
BID	Banco Interamericano de Desarrollo
CITI	Iniciativa de capacitación institucional colaborativa
DD	Diferencias en diferencias
DIME	Evaluación de impacto para el desarrollo (Banco Mundial)
DRD	Diseño de regresión discontinua
EMARF	Específico, medible, atribuible, realista y focalizado
EMD	Efecto mínimo detectable
HISP	Programa de Subsidios de Seguros de Salud
ID	Número de identificación
IDU	Instituto para el Desarrollo de Ultramar
IHSN	International Household Survey Network
ITT	Intención de tratar
IV	Variables instrumentales
J-PAL	Abdul Latif Jameel Poverty Action Lab
JRI	Junta de revisión institucional
LATE	Efecto promedio local del tratamiento
NIH	National Institutes of Health (Estados Unidos)
OMS	Organización Mundial de la Salud
ONG	Organización no gubernamental
OSF	Open Science Framework
RCT	Ensayo controlado aleatorio

RIDIE	Registry for International Development Impact Evaluations
SIEF	Fondo Estratégico de Evaluación de Impacto (Banco Mundial)
SUTVA	Supuesto de estabilidad del valor de la unidad de tratamiento
TOT	Tratamiento en los tratados
USAID	Agencia de Estados Unidos para el Desarrollo Internacional

Primera parte

INTRODUCCIÓN A LA EVALUACIÓN DE IMPACTO

La primera parte de este libro presenta una visión general de la evaluación de impacto. En el capítulo 1 se analiza por qué la evaluación de impacto es importante y cómo se inscribe en el contexto de una formulación ética de las políticas basada en evidencia. Se compara la evaluación de impacto con el monitoreo, se describen las características que definen la evaluación de impacto y se abordan enfoques complementarios, entre ellos el análisis de costo-beneficio y de costo-efectividad. Asimismo, aquí se introduce un elemento clave del libro, a saber: cómo los recursos disponibles de un programa, los criterios de elegibilidad para seleccionar a los beneficiarios y los plazos para la implementación contribuyen a definir las opciones en la selección de los métodos de evaluación de impacto. Por último, se detallan diferentes modalidades de evaluación de impacto, como la evaluación prospectiva y retrospectiva y las pruebas de eficacia frente a las pruebas de efectividad, y se concluye con un debate sobre cuándo utilizar las evaluaciones de impacto.

El capítulo 2 versa sobre cómo formular preguntas e hipótesis de evaluación que son útiles para las políticas. Estas preguntas e hipótesis forman la base de la evaluación porque definen su foco. También se expone el concepto fundamental de una teoría del cambio y los usos correspondientes de las cadenas de resultados y de los indicadores de desempeño. Este capítulo presenta la primera introducción al estudio de casos ficcionales, el Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés, *Health Insurance Subsidy Program*) utilizado a lo largo del libro y en el material de apoyo que se halla en el sitio web de evaluación de impacto en la práctica (http://www.worldbank.org/ieinpractice).

CAPÍTULO 1

¿Por qué evaluar?

La formulación de políticas basada en evidencia

Los programas y políticas de desarrollo suelen estar diseñados para cambiar resultados, como aumentar los ingresos, mejorar el aprendizaje o reducir las enfermedades. Saber si estos cambios se logran o no es una pregunta crucial para las políticas públicas, aunque a menudo no suele abordarse. Es más habitual que los administradores de los programas y los responsables de las políticas se centren en medir e informar sobre los insumos y los productos inmediatos de un programa (cuánto dinero se gasta, cuántos libros de texto se distribuyen, cuántas personas participan en un programa de empleo) en lugar de evaluar si los programas han logrado sus objetivos de mejorar los resultados.

Las evaluaciones de impacto forman parte de una agenda mas amplia de *formulación de políticas públicas basadas en evidencia*. Esta tendencia mundial creciente se caracteriza por un cambio de enfoque, ya que en lugar de centrarse en los insumos lo hace en los productos y resultados, y está reconfigurando las políticas públicas. Centrarse en los resultados no solo sirve para definir y hacer un seguimiento de los objetivos nacionales e internacionales, sino que –además– los administradores de programas utilizan y necesitan cada vez más los resultados para mejorar la rendición de cuentas, definir las asignaciones presupuestarias y orientar el diseño del programa y las decisiones de políticas.

Concepto clave

Las evaluaciones son valoraciones periódicas y objetivas de un proyecto, programa o política planificada, en curso o terminada. Las evaluaciones se utilizan para responder a preguntas específicas, a menudo relacionadas con el diseño, la implementación y/o los resultados.

El monitoreo y la evaluación son fundamentales en la formulación de políticas basadas en evidencia. Ofrecen un conjunto central de instrumentos que las partes interesadas pueden utilizar para verificar y mejorar la calidad, eficiencia y efectividad de las políticas y de los programas en diferentes etapas de implementación o, en otras palabras, para centrarse en los resultados. A nivel de la gestión del programa, es necesario saber cuáles son las opciones de diseño costo-efectivas, o demostrar ante los responsables de la toma de decisiones que los programas están logrando sus resultados previstos con el fin de obtener asignaciones presupuestarias para continuarlos o ampliarlos. A nivel nacional, los ministerios compiten unos con otros para obtener financiamiento del ministerio de Finanzas. Y, por último, los gobiernos deben rendir cuentas ante los ciudadanos para informales del resultado de los programas públicos. La evidencia puede constituir una base sólida para la transparencia y la rendición de cuentas.

La evidencia robusta generada por las evaluaciones de impacto está sirviendo cada vez más como fundamento para una mayor rendición de cuentas, innovación y aprendizaje. En un contexto en que los responsables de las políticas y la sociedad civil exigen resultados y la rendición de cuentas de los programas públicos, la evaluación de impacto puede proporcionar evidencia robusta y creíble sobre el desempeño y ante todo sobre si un programa concreto ha alcanzado o está alcanzando sus resultados deseados. Las evaluaciones de impacto también son cada vez más utilizadas para probar innovaciones en el diseño de programas o en la prestación de servicios. A nivel mundial, estas evaluaciones son fundamentales para construir conocimientos acerca de la efectividad de los programas de desarrollo, iluminando sobre lo que funciona y no funciona para reducir la pobreza y mejorar el bienestar.

En pocas palabras, una evaluación de impacto mide los cambios en el bienestar de los individuos que se pueden atribuir a un proyecto, un programa o una política específicos. Este enfoque en la atribución es el sello distintivo de las evaluaciones de impacto. Por lo tanto, el reto fundamental en una evaluación de esta naturaleza consiste en identificar la relación causal entre el programa o la política y los resultados de interés.

Las evaluaciones de impacto suelen medir el impacto promedio de un programa, las modalidades del programa o una innovación en el diseño. Por ejemplo, ¿el programa de agua y saneamiento aumentó el acceso a agua potable y mejoró los resultados de salud? ¿Un programa de estudios alternativo mejoró las puntuaciones de las pruebas de los alumnos? ¿La innovación de incluir destrezas cognitivas como parte de un programa de formación de jóvenes ha tenido éxito promoviendo la iniciativa empresarial e incrementando los ingresos? En cada uno de estos casos, la evaluación de impacto

proporciona información sobre si el programa provocó los cambios deseados en los resultados, al compararse con estudios de casos o anécdotas específicas, que solo pueden brindar información parcial y que quizá no sean representativos de los impactos generales del programa. En este sentido, las evaluaciones de impacto bien diseñadas y bien implementadas son capaces de proporcionar evidencia convincente y exhaustiva que puede ser utilizada para fundamentar las decisiones de las políticas, influir en la opinión pública y mejorar el funcionamiento de los programas.

Las evaluaciones de impacto clásicas abordan la efectividad de un programa en comparación con la ausencia del mismo. El recuadro 1.1 se refiere a la evaluación de impacto bien conocida del programa de transferencias condicionadas en México, e ilustra cómo la evaluación contribuyó a los debates de las políticas públicas en relación con la ampliación del programa.[1]

Recuadro 1.1: Cómo una evaluación exitosa puede promover la sostenibilidad política de un programa de desarrollo

El programa de transferencias condicionadas de México

En los años noventa, el gobierno de México lanzó un programa innovador de transferencias condicionadas, llamado inicialmente Progresa (que luego pasó a llamarse Oportunidades y más tarde Prospera, al tiempo que cambiaron unos cuantos elementos del mismo). Sus objetivos consistían en proporcionar a los hogares pobres un apoyo de corto plazo a los ingresos y en incentivar la inversión en el capital humano de los niños, mediante transferencias de efectivo a las madres de hogares pobres con la condición de que sus hijos asistieran a la escuela y visitaran regularmente un centro de salud.

Desde el comienzo, el gobierno consideró esencial monitorear y evaluar el programa. Los funcionarios responsables contrataron a un grupo de investigadores para que diseñaran una evaluación de impacto y la incorporaran en la ampliación del programa al mismo tiempo que este se ponía en marcha de manera progresiva en las comunidades participantes.

Las elecciones presidenciales de 2000 se saldaron con un cambio en el partido gobernante. En 2001 los evaluadores externos de Progresa presentaron su conclusiones al gobierno recién elegido. Los resultados del programa eran impresionantes: demostraban que el programa estaba bien focalizado en los pobres y que había generado cambios prometedores en el capital humano de los hogares. Schultz (2004) encontró que el programa mejoraba de forma significativa la matriculación escolar, en un promedio de 0,7 años adicionales de escolarización. Gertler (2004) observó que la incidencia de las enfermedades de los niños disminuía en un 23%, mientras que el número de días perdidos por enfermedad o

Continúa en la página siguiente.

Recuadro 1.1: Cómo una evaluación exitosa puede promover la sostenibilidad política de un programa de desarrollo *(continúa)*

discapacidad decrecía en un 19% entre los adultos. Entre los resultados nutricionales, Behrman y Hoddinott (2001) hallaron que el programa reducía la probabilidad de retraso en el crecimiento en alrededor de 1 centímetro al año en los niños durante la edad crítica de 12-36 meses.

Estos resultados de la evaluación fueron el punto de partida de un diálogo sobre las políticas basadas en evidencia y contribuyeron a la decisión del nuevo gobierno de seguir con el programa. El gobierno amplió su alcance e introdujo las becas en la enseñanza media-superior y mejoró los programas de salud para los adolescentes. Al mismo tiempo, los resultados fueron utilizados para modificar otros programas de ayuda social, como el subsidio de la tortilla, muy generoso pero no tan bien focalizado, cuya escala se redujo.

La exitosa evaluación de Progresa también contribuyó a la rápida adopción de sistemas de transferencias condicionadas en todo el mundo, y a la adopción en México de una legislación que establece la evaluación de todos los proyectos sociales.

Fuentes: Behrman y Hoddinott (2001); Fiszbein y Schady (2009); Gertler (2004); Levy y Rodríguez (2005); Schultz (2004); Skoufias y McClafferty (2001).

El recuadro 1.2 ilustra cómo la evaluación de impacto influyó en la política educativa de Mozambique, al demostrar que el preescolar comunitario puede ser una fórmula asequible y efectiva de abordar la educación temprana y promover la matriculación de los niños en la escuela primaria a la edad adecuada.

Además de abordar la pregunta básica de si un programa es efectivo o no, las evaluaciones de impacto también se pueden utilizar para probar explícitamente modalidades de programas alternativos o innovaciones de diseño. A medida que los responsables de las políticas se centran cada vez más en entender mejor cómo perfeccionar la implementación y obtener más a cambio del dinero, los enfoques que prueban alternativas de diseño están ganando terreno rápidamente. Por ejemplo, una evaluación puede comparar el desempeño de un programa de formación con el de una campaña promocional para ver cuál es más efectivo para aumentar la alfabetización financiera. Una evaluación de impacto puede probar cuál es el enfoque de la combinación de nutrición y de estimulación del niño que tiene el mayor impacto en su desarrollo. O la evaluación puede probar una innovación de diseño para mejorar el diseño de un programa existente, como utilizar mensajes de texto para mejorar el cumplimiento cuando se trata de tomar la medicación prescrita.

Recuadro 1.2: El impacto de las políticas de un modelo preescolar innovador

Desarrollo preescolar y de la primera infancia en Mozambique

Si bien el preescolar se reconoce como una buena inversión y un enfoque efectivo para preparar a los niños para la escuela y las etapas posteriores de la vida, los países en desarrollo se han enfrentado a la pregunta de cómo introducir un modelo de preescolar escalable y costo-efectivo. En Mozambique solo alrededor del 4% de los niños asiste a preescolar. Al llegar a la escuela primaria, algunos niños de comunidades rurales muestran señales de retrasos en el desarrollo y a menudo no están preparados para las demandas de la escuela primaria. Además, a pesar de que en la escuela primaria hay una tasa de matriculación de casi el 95%, a una tercera parte de los niños no se los inscribe a la edad adecuada.

En 2006 Save the Children lanzó un programa piloto de preescolar comunitario en zonas rurales de Mozambique con la intención de mejorar el desarrollo cognitivo, social, emocional y físico de los niños. En lo que se considera la primera evaluación aleatorizada de un programa preescolar en África rural, en 2008 un equipo de investigación llevó a cabo una evaluación de impacto del programa. Sobre la base de los resultados positivos de la evaluación, el gobierno de Mozambique adoptó el modelo preescolar comunitario de Save the Children y decidió ampliarlo a 600 comunidades.

La evaluación llegó a la conclusión de que los niños que asistían a preescolar tenían un 24% más de probabilidades de matricularse en la escuela primaria que los niños del grupo de comparación, y un 10% más de probabilidades de comenzar a la edad adecuada. En la escuela primaria, los niños que habían asistido a preescolar dedicaron casi un 50% más de tiempo a las tareas escolares y a otras actividades relacionadas con la escuela que los que no habían asistido. La evaluación también mostraba aumentos positivos en la preparación para la escuela; los niños que habían cursado preescolar obtenían mejores resultados en las pruebas cognitivas y socioemocionales, y alcanzaban un buen desarrollo motor versus el grupo de comparación.

Otros miembros del hogar también se beneficiaron de la matriculación de los niños en preescolar al disponer de más tiempo para dedicar a actividades productivas. Cuando en un hogar había un niño que concurría a preescolar, los hermanos mayores tenían un 6% más de probabilidades de asistir a la escuela y los cuidadores tenían un 26% más de probabilidades de haber trabajado en los últimos 30 días.

Esta evaluación demostró que incluso en un entorno de bajos ingresos, el preescolar puede ser una manera efectiva de promover el desarrollo cognitivo, preparar a los niños para la escuela primaria y aumentar la probabilidad de que comiencen la escuela primaria a la edad adecuada.

Fuente: Martínez, Nadeau y Pereira (2012).

¿Qué es la evaluación de impacto?

La evaluación de impacto es uno de los numerosos métodos que existen para apoyar las políticas públicas basadas en evidencia, incluidos el monitoreo y otros tipos de evaluación.

El *monitoreo* es un proceso continuo mediante el cual se lleva a cabo un seguimiento de lo que ocurre con un programa y se utilizan los datos recopilados para fundamentar la implementación de los programas y la administración y las decisiones diarias. A partir sobre todo de datos administrativos, el monitoreo realiza un seguimiento de los desembolsos financieros y del desempeño del programa en relación con los resultados esperados, y analiza las tendencias a lo largo del tiempo.[2] El monitoreo es necesario en todos los programas y constituye una fuente crítica de información sobre el desempeño de los mismos, lo cual abarca también la implementación y los costos. Normalmente, el monitoreo se aplica a los insumos, actividades y productos, aunque ocasionalmente también puede abarcar los resultados, como, por ejemplo, el progreso alcanzado en los objetivos de desarrollo nacional.

Las *evaluaciones* son valoraciones periódicas y objetivas de un proyecto, programa o política planificado, en curso o terminado. Se utilizan para responder a preguntas específicas relacionadas con el diseño, la implementación y los resultados. En contraste con el monitoreo, que es permanente, las evaluaciones se llevan a cabo en momentos concretos en el tiempo y a menudo requieren una perspectiva externa de los técnicos expertos. Su diseño, método y costo varían considerablemente en función del tipo de pregunta que la evaluación intente responder. En términos generales, las evaluaciones pueden abordar tres tipos de preguntas (Imas y Rist, 2009):[3]

- *Preguntas descriptivas*, que apuntan a lo que está ocurriendo. Se centran en los procesos, las condiciones, las relaciones organizacionales y las opiniones de las partes interesadas.
- *Preguntas normativas,* que comparan lo que ocurre con lo que debería ocurrir. Evalúan las actividades e investigan si los objetivos se cumplen o no. Estas preguntas pueden aplicarse a los insumos, las actividades y los productos.
- *Preguntas de causa y efecto,* que se centran en la atribución. Investigan qué diferencia produce la intervención en los resultados.

Hay numerosos tipos de evaluación y de métodos de evaluación, basados en datos cuantitativos y cualitativos. Los *datos cualitativos* no se expresan en números sino más bien mediante un lenguaje o, a veces, imágenes. Los *datos cuantitativos* son mediciones numéricas y habitualmente se asocian con escalas o métricas. Tanto los unos como los otros se pueden utilizar para responder al tipo de preguntas planteado más arriba. En la práctica, numerosas evaluaciones trabajan con ambos tipos de datos. Hay múltiples fuentes de datos que se pueden emplear en las evaluaciones, tanto datos primarios recopilados para el objetivo de la evaluación como los datos secundarios disponibles (véase el capítulo 16 sobre las fuentes de datos).

Este libro se centra en las evaluaciones de impacto que se sirven de datos cuantitativos pero subrayan el valor del monitoreo, de los métodos de evaluación complementarios y del uso tanto de datos cuantitativos como cualitativos.

Las *evaluaciones de impacto* constituyen un tipo particular de evaluación que pretende responder a una pregunta específica de causa y efecto: ¿Cuál es el impacto (o efecto causal) de un programa en un resultado de interés? Esta pregunta básica incorpora una dimensión causal importante. Se centra únicamente en el *impacto*, es decir, en los cambios *directamente atribuibles* a un programa, una modalidad de programa o una innovación de diseño.

La pregunta básica de la evaluación –¿cuál es el impacto o efecto causal de un programa en un resultado de interés?– se puede aplicar en numerosos contextos. Por ejemplo, ¿cuál es el efecto causal de las becas en la asistencia escolar y los logros académicos? ¿Cuál es el impacto en el acceso a la atención sanitaria de contratar la atención primaria con proveedores privados? Si los suelos de tierra son reemplazados por suelos de cemento, ¿cuál será el impacto en la salud de los niños? ¿La mejora de los caminos aumenta el acceso a los mercados laborales e incrementa el ingreso de los hogares y, en caso afirmativo, en qué medida? ¿Influye el tamaño de la clase en los logros de los alumnos y, en caso afirmativo, en qué medida? Como muestran estos ejemplos, la pregunta de la evaluación básica se puede ampliar para analizar el impacto de una *modalidad de programa o innovación de diseño*, no solo de un programa.

Concepto clave

Las evaluaciones de impacto pretenden responder un tipo particular de pregunta: ¿Cuál es el impacto (o efecto causal) de un programa en un resultado de interés?

El sello distintivo de las evaluaciones de impacto es centrarse en la causalidad y la atribución. Todos los métodos de evaluación de impacto plantean alguna forma de pregunta de *causa y efecto*. El enfoque para abordar la causalidad determina las metodologías que se pueden utilizar. Para estimar el efecto causal o el impacto de un programa en los resultados, cualquier método de evaluación de impacto elegido debe estimar el llamado *contrafactual*, es decir: cuál habría sido el resultado de los participantes del programa si no hubieran participado en el mismo. En la práctica, la evaluación de impacto requiere que el equipo de evaluación encuentre un grupo de comparación para estimar qué les habría ocurrido a los participantes del programa sin el programa, y luego efectuar comparaciones con el grupo de tratamiento que ha sido objeto del programa. En la segunda parte de este libro se describen los principales métodos que se pueden aplicar para encontrar grupos de comparación adecuados.

Concepto clave

La elección de un método de evaluación de impacto depende de las características operativas del programa que se evalúa, sobre todo de sus recursos disponibles, sus criterios de elegibilidad para seleccionar a los beneficiarios y los plazos para la implementación del programa.

Uno de los principales mensajes de este libro es que la elección de un método de evaluación de impacto depende de las características operativas del programa que se evalúa. Cuando las reglas de operación del programa son equitativas y transparentes y contemplan la rendición de cuentas, siempre se podrá encontrar un buen diseño de evaluación de impacto, ya sea que

la evaluación de impacto se planifique al comienzo, o durante el proceso de diseño o de implementación de un programa. El contar con reglas de operación claras y bien definidas para un programa no solo tiene un valor intrínseco en las políticas públicas y en una gestión solvente de los programas: también es esencial para construir buenos grupos de comparación, lo cual constituye la base de las evaluaciones de impacto rigurosas. Concretamente, la elección de un método de evaluación de impacto está determinada por las características operativas del programa, en particular sus recursos disponibles, los criterios de elegibilidad para seleccionar a los beneficiarios y los plazos para la implementación del programa. Como se verá en las partes 2 y 3 de este libro, se pueden formular tres preguntas acerca del contexto operativo de un determinado programa: ¿El programa tiene recursos para servir a todos los beneficiarios elegibles? ¿El programa está focalizado o es universal? ¿El programa se ofrecerá a todos los beneficiarios de una sola vez o de manera secuencial? La respuesta a estas tres preguntas determinará cuál de los métodos presentados en la parte 2 –asignación aleatoria, variables instrumentales, regresión discontinua, diferencias en diferencias o pareamiento– es el más adecuado para un determinado contexto operativo.

Evaluación de impacto prospectiva versus evaluación retrospectiva

Las evaluaciones de impacto se pueden dividir en dos categorías: prospectivas y retrospectivas. Las *evaluaciones prospectivas* se desarrollan simultáneamente con el diseño del programa y se incorporan en la implementación del mismo. Los datos de línea de base se recopilan antes de implementar el programa, tanto en el grupo que recibe la intervención (denominado *grupo de tratamiento*) como en el grupo utilizado como comparación y que no es objeto de la intervención (denominado *grupo de comparación*). Las *evaluaciones retrospectivas* evalúan el impacto del programa después de que se lo haya implementado, y los grupos de tratamiento y de comparación se generan ex post.

Las evaluaciones de impacto prospectivas tienen más probabilidades de producir resultados solventes y creíbles, por tres motivos. En primer lugar, se pueden recopilar datos de línea de base para establecer las medidas de los resultados de interés antes de que el programa haya comenzado. Estos datos son importantes para medir los resultados antes de la intervención. Los datos de línea de base en los grupos de tratamiento y comparación se deben analizar para asegurar que los grupos sean similares. Las líneas de base también se pueden utilizar para evaluar la efectividad de la focalización, es decir, si un programa llega o no a sus beneficiarios previstos.

En segundo lugar, definir la medida de éxito del programa en la etapa de planificación del mismo centra tanto el programa como la evaluación en los resultados previstos. Como se verá, las evaluaciones de impacto se basan en la teoría del cambio de un programa o una cadena de resultados. El diseño de una evaluación de este tipo contribuye a clarificar los objetivos del programa, sobre todo porque requiere establecer medidas bien definidas de su éxito. Los responsables de las políticas deberían definir objetivos claros para el programa y formular preguntas claras que la evaluación debe contestar, para garantizar que los resultados sean relevantes para las políticas. En realidad, el pleno apoyo de los responsables de las políticas es un requisito necesario para el éxito de una evaluación; no se deberían emprender evaluaciones de impacto a menos que los responsables de las políticas estén convencidos de la legitimidad de las mismas y de su valor para fundamentar decisiones clave de las políticas públicas.

En tercer lugar, y lo que es aún más importante, en una evaluación prospectiva los grupos de tratamiento y comparación se definen antes de implementar la intervención que será evaluada. Como se explicará con mayor detalle en los próximos capítulos, existen muchas otras opciones para llevar a cabo evaluaciones válidas cuando las evaluaciones se planifican desde el comienzo, antes de que la implementación tenga lugar. En las partes 2 y 3 se argumenta que casi siempre es posible encontrar una estimación válida del contrafactual para cualquier programa cuyas reglas de asignación sean claras y transparentes, siempre que la evaluación se diseñe de manera prospectiva. En resumen, las evaluaciones prospectivas son las que tienen más probabilidades de generar contrafactuales válidos. En la etapa de diseño, se pueden contemplar maneras alternativas de estimar un contrafactual válido. El diseño de la evaluación de impacto también se puede alinear plenamente con las reglas operativas del programa, así como con el despliegue o el proceso de expansión de este último.

Concepto clave
Las evaluaciones prospectivas se diseñan y elaboran antes de implementar un programa.

Por el contrario, en las evaluaciones retrospectivas, el equipo que lleva a cabo la evaluación a menudo tiene información tan limitada que resulta difícil analizar si el programa fue implementado con éxito y si sus participantes realmente se beneficiaron de él. Numerosos programas no recopilan datos de línea de base a menos que se haya incorporado la evaluación desde el principio, y una vez que el programa está funcionando ya es demasiado tarde para hacerlo.

Las evaluaciones retrospectivas que utilizan los datos existentes son necesarias para evaluar los programas creados en el pasado. En estas situaciones, las opciones para obtener una estimación válida del contrafactual son mucho más limitadas. La evaluación depende de reglas claras de operación del programa en lo que respecta a la asignación de beneficios. También depende de la disponibilidad de datos con suficiente cobertura sobre los

grupos de tratamiento y comparación, tanto antes como después de la implementación del programa. El resultado es que la viabilidad de una evaluación retrospectiva depende del contexto y nunca está garantizada. Incluso cuando son viables, las evaluaciones retrospectivas a menudo utilizan métodos cuasi experimentales y dependen de supuestos más fuertes y, por ende, pueden producir evidencia más discutible.[4]

Estudios de eficacia y estudios de efectividad

La función principal de la evaluación de impacto consiste en producir evidencia sobre el desempeño de un programa a fin de que sea utilizada por los funcionarios públicos, los administradores del programa, la sociedad civil y otros actores relevantes. Los resultados de las evaluaciones de impacto son particularmente útiles cuando las conclusiones se pueden aplicar a una población de interés más amplia. La cuestión de la generalización es clave para los responsables de las políticas, puesto que determina si los resultados identificados en la evaluación pueden replicarse en grupos ajenos a los que han sido estudiados en la evaluación si aumenta la escala del programa.

En los primeros tiempos de las evaluaciones de impacto de los programas de desarrollo, una gran parte de la evidencia se basaba en *estudios de eficacia*, es decir, pruebas llevadas a cabo en un entorno específico en condiciones rigurosamente controladas para asegurar la consistencia entre el diseño de la evaluación y la implementación del programa. Dado que los estudios de eficacia suelen realizarse como experiencias piloto con una amplia participación técnica de los investigadores mientras el programa se está implementando, puede que sus resultados, a menudo de pequeña escala, no ofrezcan necesariamente mucha información acerca del impacto de un proyecto similar implementado a mayor escala en circunstancias normales. Los estudios de eficacia analizan la prueba de concepto, a menudo para sondear la viabilidad de un nuevo programa o una teoría específica del cambio. Si el programa no genera impactos anticipados bajo estas condiciones cuidadosamente manejadas, es poco probable que funcione si se despliega en circunstancias normales. Por ejemplo, una intervención piloto que introduce nuevos protocolos de tratamiento médico puede funcionar en un hospital con excelentes administradores y equipo médico, pero puede que la misma intervención no funcione en un hospital promedio con administradores menos esmerados y limitaciones de personal. Además, los cálculos de costo-beneficio variarán, dado que los pequeños estudios de eficacia quizá no capturen los costos fijos ni las economías de escala. Como consecuencia, si bien la evidencia de los estudios de eficacia puede ser útil para probar un enfoque innovador, los resultados a menudo tienen una capacidad de generalización

limitada y no siempre representan adecuadamente entornos más generales, que suelen ser la principal preocupación de los responsables de las políticas.

Al contrario, los *estudios de efectividad* proporcionan evidencia a partir de las intervenciones que tienen lugar en circunstancias normales, utilizando vías de implementación regulares y con el objeto de producir conclusiones que se pueden generalizar para una población grande. Cuando las evaluaciones de efectividad están adecuadamente diseñadas e implementadas, los resultados pueden ser generalizables para beneficiarios previstos fuera de la muestra de la evaluación, siempre y cuando la ampliación utilice las mismas estructuras de implementación y llegue a poblaciones similares a la de la muestra de la evaluación. Esta validez externa tiene una importancia crítica para los responsables de las políticas porque les permite utilizar los resultados de la evaluación para fundamentar decisiones que afectan a todo el programa y que se aplican a los beneficiarios previstos más allá de la muestra de la evaluación (véase el recuadro 1.3).

Concepto clave

Los *estudios de eficacia* evalúan si un programa *puede* funcionar en condiciones ideales, mientras que los *estudios de efectividad* evalúan si un programa *realmente* funciona en condiciones normales.

Recuadro 1.3: Pruebas de la capacidad generalizable de los resultados

Una evaluación multisitio del enfoque de "graduación" para aliviar la extrema pobreza

Al evaluar un programa en múltiples contextos, los investigadores pueden analizar si los resultados de una evaluación de impacto se pueden generalizar. Estas denominadas *evaluaciones multisitio* contribuyen al creciente corpus de evidencia sobre qué funciona y qué no lo hace en el desarrollo, y pueden proporcionar perspectivas clave a los responsables de las políticas en diferentes países.

Por ejemplo, en 2007 Banerjee et al. iniciaron una evaluación multisitio del enfoque de "graduación" para aliviar la extrema pobreza. El modelo había sido objeto de gran atención en todo el mundo después de haber arrojado resultados impresionantes en Bangladesh. Puesto en marcha por el Bangladesh Rural Advancement Committee (BRAC), una gran organización de desarrollo global, el modelo se proponía ayudar a los muy pobres a "graduarse" de la extrema pobreza mediante transferencias de efectivo, activos productivos y formación intensiva.

Banerjee y sus colegas pretendían analizar si el enfoque de graduación podía funcionar en diferentes países a través de seis evaluaciones de impacto aleatorias simultáneas en Etiopía, Ghana, Honduras, India, Pakistán y Perú. En cada país, los investigadores trabajaron con organizaciones no gubernamentales (ONG) locales para implementar un programa de graduación similar.

Continúa en la página siguiente.

Recuadro 1.3: Pruebas de la capacidad generalizable de los resultados *(continúa)*

Si bien el programa se modificó para adecuarse a los diferentes contextos en cada país, los principios clave seguían siendo los mismos. El programa se centró en los hogares más pobres en pueblos de las regiones más pobres de cada país. Durante 24 meses, los hogares beneficiarios recibieron activos productivos, formación y apoyo, *coaching* en habilidades para la vida, dinero, información sanitaria y ayuda en la inclusión financiera. La evaluación de impacto medía la efectividad de proporcionar este paquete de beneficios.

El estudio evaluó los impactos del programa en 10 conjuntos de resultados. Un año después de que el programa terminara en los seis países, se produjeron mejoras considerables en ocho de los 10 conjuntos de resultados: consumo per cápita, seguridad alimentaria, valor de los activos, inclusión financiera, tiempo dedicado a trabajar, ingresos y rentas, salud mental y participación política. La magnitud de los impactos variaba según los países, y hubo impactos considerables en el valor de los activos en todos los países excepto uno. No se registraron impactos estadísticamente significativos en el índice de salud física.

Los resultados también variaban de un país a otro. Las mejoras en el consumo per cápita no fueron significativas en Honduras ni en Perú, y la mejora en el valor de los activos no lo fue en Honduras. Sin embargo, en términos agregados, la evaluación apuntaba hacia la promesa de este tipo de intervención multifacética para mejorar las vidas de las familias muy pobres en una gama de entornos.

Fuentes: Banerjee et al. (2015); BRAC (2013).

Enfoques complementarios

Como se ha señalado, las evaluaciones de impacto responden a preguntas específicas de causa y efecto. Otros enfoques –entre ellos un estrecho *monitoreo* del programa, y también el uso complementario de otros métodos de evaluación, como *simulaciones ex ante, análisis con métodos mixtos* que se basan en datos cualitativos y cuantitativos, y *evaluaciones de procesos*– pueden servir como valiosos complementos de las evaluaciones de impacto. Estos otros enfoques tienen numerosas aplicaciones útiles, como estimar el efecto de las reformas antes de que sean implementadas, contribuir a focalizar las preguntas centrales de la evaluación de impacto, realizar seguimientos de la implementación del programa e interpretar los resultados de las evaluaciones de impacto.

Las evaluaciones de impacto que se realizan en aislamiento con respecto a otras fuentes de información son vulnerables en términos tanto de su calidad técnica como de su relevancia para las políticas públicas. Si bien los resultados de dichas evaluaciones pueden proporcionar evidencia robusta para saber si ha tenido lugar un efecto, a menudo existen limitaciones para proporcionar una perspectiva clara de los canales a través de los cuales la política o programa influyó en los resultados observados. Sin información de

las evaluaciones de procesos sobre la naturaleza y el contenido del programa para contextualizar los resultados de la evaluación, puede que los responsables de las políticas queden confundidos acerca de por qué se alcanzaron o no ciertos resultados. Además, sin datos de monitoreo sobre cómo, cuándo y dónde se está implementando el programa, la evaluación será ciega en cuanto a si los beneficios llegaron a los beneficiarios previstos y cuándo lo hicieron, o bien si alcanzaron de forma inintencionada al grupo de comparación.

El monitoreo

El monitoreo de la implementación del programa, las más de las veces mediante el uso de datos administrativos, es crítico en una evaluación de impacto. Permite al equipo de evaluación verificar si las actividades se están realizando según lo planificado, es decir, a qué participantes se les adjudicó el programa, con qué rapidez se amplió este último, y cómo se están gastando los recursos. Esta información es fundamental para implementar la evaluación, por ejemplo, para asegurar que los datos de línea de base se recopilen antes de que se introduzca el programa en la muestra de la evaluación y para verificar la integridad de los grupos de tratamiento y comparación. El monitoreo es esencial para verificar si un beneficiario realmente participa en el programa y para que no intervengan los no beneficiarios. Además, los datos administrativos pueden proporcionar información sobre el costo de implementación del programa, lo cual también es necesario para los análisis de costo-beneficio y costo-efectividad.

Simulaciones ex ante

Las *simulaciones ex ante* son evaluaciones que utilizan datos disponibles para simular los efectos esperados de una reforma de programas o políticas en los resultados de interés. Pueden ser muy útiles para medir la efectividad esperada relativa de una gama de opciones de diseño de programas alternativos en los resultados. Se trata de métodos habitualmente usados que dependen de la disponibilidad de datos de gran alcance y calidad que se pueden utilizar para aplicar modelos de simulación adecuados a la pregunta en cuestión (véase el recuadro 1.4). Al contrario de las evaluaciones de impacto, estos métodos se emplean para simular futuros efectos potenciales, más que para medir los impactos reales de los programas implementados. Este tipo de métodos puede ser sumamente útil para establecer referencias para los probables efectos del programa y para instituir objetivos realistas, así como para estimar costos, tasas de retorno y otros parámetros económicos. Se suelen utilizar como la base de los análisis económicos de los proyectos, especialmente antes de que se introduzca una reforma o se implemente un proyecto.

Recuadro 1.4: Simulación de posibles efectos del proyecto a través del modelado estructural

Construcción de un modelo para probar diseños alternativos utilizando datos de Progresa en México

Se puede utilizar un cierto tipo de simulación ex ante (*modelado estructural*) para estimar los efectos de un programa en una gama de diseños alternativos. En la evaluación Progresa/Oportunidades/Prospera, que se describe en el recuadro 1.1, los datos recopilados eran lo suficientemente ricos para que los investigadores construyeran un modelo que podía simular los efectos esperados de diseños de programas alternativos.

Todd y Wolpin (2006) utilizaron datos de línea de base de la evaluación de impacto para construir un modelo de las decisiones de los padres a propósito de sus hijos, incluida la escolarización. Los autores simularon cómo serían los efectos con distintos diseños de programa, y descubrieron que si el programa eliminaba los incentivos en efectivo para la asistencia escolar en los primeros años y, en su lugar, utilizaba el dinero para aumentar los incentivos en efectivo para los alumnos de cursos superiores, los efectos en la escolarización promedio completada probablemente serían mayores.

En este caso, las proyecciones se realizaron utilizando la encuesta de línea de base de una evaluación de impacto que ya se había realizado. Los resultados de las predicciones se pudieron probar para ver si arrojaban los mismos impactos que el experimento del programa real. Sin embargo, esto no es posible de hacer normalmente. Este tipo de métodos de simulación suele utilizarse antes de que el programa realmente se implemente con el fin de analizar los probables efectos de diversos diseños de programa alternativos. Así, pueden proporcionar una base para estrechar la gama de opciones a probarse en la práctica.

Fuente: Todd y Wolpin (2006).
Nota: Para otro ejemplo de modelado estructural, véase Bourguignon, Ferreira y Leite (2003).

Los métodos mixtos

Los enfoques de *métodos mixtos* que combinan datos cuantitativos y cualitativos constituyen un complemento clave en las evaluaciones de impacto que se basan únicamente en el uso de datos cuantitativos, sobre todo para contribuir a generar hipótesis y enfocar las preguntas de la investigación antes de recopilar los datos cuantitativos, así como para presentar perspectivas y visiones novedosas del desempeño de un programa durante y después de su implementación. Hay numerosos métodos cualitativos, que componen su propio ámbito de investigación.[5] Los métodos que generan datos cualitativos suelen basarse en enfoques abiertos, que no dependen de las respuestas predeterminadas de las personas entrevistadas. Los datos se generan a través de una gama de enfoques, incluidos grupos focales, historiales y entrevistas con beneficiarios seleccionados y otros informantes clave (Rao y Woolcock, 2003). También pueden

incluir una gama de evaluaciones observacionales y etnográficas. A pesar de que las observaciones, ideas y opiniones recopiladas durante el trabajo cualitativo no suelen ser estadísticamente representativas de los beneficiarios del programa –y, por lo tanto, no son generalizables– resultan útiles para entender por qué se han alcanzado o no ciertos resultados (recuadro 1.5).

Las evaluaciones que integran el análisis cuantitativo y cualitativo se caracterizan por utilizar *métodos mixtos* (Bamberger, Rao y Woolcock, 2010). En el

Recuadro 1.5: Un método mixto de evaluación en acción

Combinación de una prueba controlada aleatoria con un estudio etnográfico en India

Los enfoques de métodos mixtos pueden ser especialmente útiles cuando evalúan programas con resultados que son difíciles de medir en las encuestas cuantitativas. Los programas de los ámbitos de democracia y gobernanza constituyen ejemplos de este tipo.

Así, mientras se diseñaba una estrategia de evaluación para el programa "Campaña del pueblo" que pretendía mejorar la participación ciudadana en los gobiernos locales, Ananthpur, Malik y Rao (2014) integraron un ensayo controlado aleatorio (RCT, por sus siglas en inglés, *Randomized Control Trial*) (véase el glosario) con un estudio etnográfico llevado a cabo en un subconjunto del 10% de la muestra de evaluación utilizada para el RCT. Se emplearon métodos de pareamiento para asegurar características similares entre pueblos de tratamiento y de comparación en la muestra para el estudio cualitativo. Se asignó un experimentado investigador de campo para que viviera en cada pueblo y estudiara los impactos del programa en las estructuras sociales y políticas del pueblo.

El estudio etnográfico continuó durante dos años después de que terminó el RCT, lo que permitió observaciones de efectos a más largo plazo. Si bien el RCT encontró que la intervención no tenía un impacto estadístico significativo, el estudio cualitativo proporcionó visiones novedosas de las causas del "fracaso" de la intervención. La investigación cualitativa identificó diversos factores que obstaculizaron la efectividad de la esta última: las variaciones en la calidad de la facilitación del programa, la falta de apoyo de arriba hacia abajo y las arraigadas estructuras de poder local.

La evidencia cualitativa también descubrió algunos impactos del programa menos tangibles e inesperados. En los pueblos del tratamiento, el programa mejoró la resolución de conflictos en la prestación de servicios y aumentó la participación de las mujeres en las actividades de desarrollo de sus comunidades. Además, los investigadores de campo observaron que los gobiernos locales funcionaban mejor en los pueblos de tratamiento.

Sin la comprensión matizada del contexto y de la dinámica local que proporciona el componente cualitativo, los investigadores no habrían podido entender por qué los datos cuantitativos no encontraron impactos. El estudio etnográfico fue capaz de proporcionar una evaluación más rica, con perspectivas novedosas de los elementos útiles para mejorar el programa.

Fuente: Ananthpur, Malik y Rao (2014).

desarrollo de un enfoque de método mixto, Creswell (2014) define tres aproximaciones básicas:

1. *Convergente paralelo*. Se recopilan simultáneamente datos cuantitativos y cualitativos y se utilizan para triangular los hallazgos o para generar los primeros resultados sobre cómo se está implementando el programa y cómo lo perciben los beneficiarios.

2. *Explicativo secuencial*. Los datos cualitativos proporcionan contexto y explicaciones para los resultados cuantitativos, para explorar casos "atípicos" de éxito y fracaso, y para desarrollar explicaciones sistemáticas del desempeño del programa, como se constató en los resultados cuantitativos. De esta manera, el trabajo cualitativo puede contribuir a determinar por qué en el análisis cuantitativo se observan ciertos resultados, y se pueden usar para entrar en la "caja negra" de lo que ocurrió en el programa (Bamberger, Rao y Woolcock, 2010).

3. *Exploratorio secuencial*. El equipo de evaluación puede utilizar grupos focales, listas, entrevistas con informantes clave y otros enfoques cualitativos para desarrollar hipótesis a propósito de cómo y por qué el programa funcionaría, y para clarificar preguntas acerca de la investigación que hay que abordar en el trabajo cuantitativo de evaluación de impacto, lo que incluye las alternativas más relevantes del diseño de programas que deben ser probadas a través de la evaluación de impacto.

Las evaluaciones de procesos

Las *evaluaciones de procesos* se centran en cómo se implementa y funciona un programa, considerando si corresponde a su diseño original, y documentando su desarrollo y funcionamiento. Normalmente, estas evaluaciones pueden llevarse a cabo con relativa rapidez y a un costo razonable. En los proyectos piloto y en las etapas iniciales de un programa, pueden ser una valiosa fuente de información sobre cómo mejorar la implementación del programa, y se suelen utilizar como primeros pasos para desarrollar un programa de modo que los ajustes operativos se puedan hacer antes de que se termine su diseño. Pueden probar si un programa funciona como estaba diseñado y si es consistente con la teoría del cambio del mismo (recuadro 1.6).

Una evaluación de procesos debería incluir los siguientes elementos, que a menudo se basan en una cadena de resultados o modelo lógico (véase el capítulo 2), complementados con documentos del programa y entrevistas con informantes clave y grupos focales beneficiarios:[6]

- Objetivos del programa y contexto en el que funciona.
- Descripción del proceso utilizado para diseñar e implementar el programa.

Recuadro 1.6: Fundamentos para una ampliación a escala nacional mediante una evaluación de procesos en Tanzania

En el desempeño de un programa hay múltiples facetas. La evidencia de las evaluaciones de procesos puede complementar los resultados de la evaluación de impacto y proporcionar un cuadro más completo de dicho desempeño. Esto puede ser particularmente importante para que los programas piloto arrojen luz sobre cómo están funcionando las nuevas instituciones y los nuevos procesos.

En 2010 el gobierno de Tanzania decidió llevar a cabo en tres distritos un plan piloto de transferencias condicionadas con base en la comunidad. El programa proporcionaba una transferencia de efectivo a los hogares pobres en función del cumplimiento de ciertos requisitos educativos y sanitarios. Los grupos comunitarios ayudaron a asignar las transferencias a los hogares más vulnerables de sus comunidades. Para evaluar si este sistema basado en la comunidad funcionaba en el contexto de Tanzania, un equipo de investigadores del Banco Mundial decidió integrar una evaluación de procesos en una evaluación de impacto tradicional.

Para la evaluación de procesos se utilizaron datos cualitativos y cuantitativos. Un año después de implementar la encuesta de línea de base en distritos piloto, los investigadores organizaron un ejercicio de tarjetas de puntuación en la comunidad para calificar aspectos del programa, basándose en grupos focales compuestos por miembros de la comunidad. Estos grupos también se usaron para dar lugar a minuciosas discusiones sobre los impactos del programa que podrían ser difíciles de cuantificar, como los cambios en las relaciones entre los miembros del hogar o la dinámica de la comunidad. El objetivo de la evaluación del proceso consistía en entender cómo funcionaba el programa en la práctica y presentar recomendaciones de mejoras.

La evaluación de impacto descubrió que el programa tenía impactos positivos y estadísticamente significativos en resultados clave de educación y salud. Los niños de los hogares que participaban tenían un 15% más de probabilidades de completar la escuela primaria y un 11% menos de probabilidades de caer enfermos. Además, los grupos focales con los maestros revelaron que los alumnos de los grupos de tratamiento estaban más preparados y más atentos.

Sin embargo, los grupos focales con miembros de la comunidad indicaban que había cierto grado de descontento con el proceso de selección de los beneficiarios. Los participantes se quejaban de falta de transparencia en la selección y de retrasos en los pagos. La evaluación del proceso permitió a los administradores del programa abordar estos problemas y mejorar el funcionamiento del programa.

El trabajo de evaluación fundamentó la decisión del gobierno de Tanzania de aumentar la escala del programa. Se espera que las transferencias condicionadas con base en las comunidades lleguen a casi un millón de hogares hacia 2017, teniendo en cuenta las lecciones de esta evaluación exhaustiva.

Fuentes: Berman (2014); Evans et al. (2014).

- Descripción de las operaciones del programa, incluido cualquier cambio en las mismas.
- Datos básicos sobre las operaciones del programa, incluidos indicadores financieros y de cobertura.
- Identificación y descripción de eventos que escapan al control del programa que pueden haber influido en la implementación y los resultados.
- Documentación, como notas de concepto, manuales operativos, actas de las reuniones, informes y memorandos.

Aplicar una evaluación de impacto a un programa cuyos procesos operativos no han sido validados plantea el doble riesgo de que se malgasten los recursos de dicha evaluación, cuando en realidad podría bastar con una evaluación de proceso más sencilla, o bien el riesgo de que los ajustes necesarios en el diseño del programa se introduzcan una vez que la evaluación de impacto ya ha comenzado, lo cual cambia el carácter del programa que se evalúa y la utilidad de la mencionada evaluación.

Análisis de costo-beneficio y costo-efectividad

Conceptos clave

El análisis de costo-beneficio estima los beneficios totales esperados de un programa, comparado con sus costos totales esperados.

El análisis de costo-efectividad compara el costo relativo de dos o más programas o de alternativas de programas para alcanzar un resultado común.

Es sumamente importante que la evaluación de impacto pueda complementarse con información sobre el costo del proyecto, del programa o de la política que se evalúa.

Una vez que están disponibles los resultados de la evaluación de impacto, estos pueden combinarse con información sobre los costos del programa para responder a otras dos preguntas. En primer lugar, en la forma básica de una evaluación de impacto, añadir información del costo permitirá llevar a cabo un análisis de costo-beneficio, a partir de lo cual se podrá responder a la pregunta: ¿cuáles son los beneficios de un programa con un determinado costo? El *análisis de costo-beneficio* estima los beneficios totales esperados de un programa, comparado con sus costos totales esperados. Busca cuantificar todos los costos y beneficios de un programa en términos monetarios, y evalúa si estos últimos superan a los costos.[7]

En un mundo ideal, el análisis de costo basado en la evidencia de la evaluación de impacto existiría no solo para un programa concreto sino también para una serie de programas o alternativas de programas, de modo que los responsables de las políticas pudieran valorar qué programa o alternativa es más efectivo en función de los costos para lograr un determinado objetivo. Cuando una evaluación de impacto ensaya alternativas de un programa, añadir información sobre costos le permite responder a la segunda pregunta: ¿cómo se comparan diversas alternativas de implementación en cuanto a su costo-efectividad? Este *análisis de costo-efectividad* compara el

costo relativo de dos o más programas o alternativas de programas para alcanzar un objetivo común, como la producción agrícola o las puntuaciones de los alumnos en las pruebas.

En un análisis de costo-beneficio o de costo-efectividad, la evaluación de impacto estima el lado del beneficio o el lado de la efectividad, mientras que examinar los costos proporciona la información sobre los mismos. Este libro se centra en la evaluación de impacto y no aborda en detalle cómo recopilar datos sobre costos o llevar a cabo análisis de costo-beneficio o costo-efectividad. Sin embargo, es fundamental que la evaluación de impacto se complemente con información sobre el costo del proyecto, del programa o de la política que se evalúa. Una vez que se disponga de información sobre el impacto y el costo de diversos programas, el análisis de costo-efectividad puede identificar cuáles son las inversiones que arrojan la tasa más alta de retorno y permiten a los responsables de las políticas tomar decisiones bien fundamentadas sobre las intervenciones en las que hay que invertir. El recuadro 1.7 ilustra cómo se pueden utilizar las evaluaciones de impacto para identificar los programas más efectivos en términos de costo-efectividad y mejorar la asignación de recursos.

Recuadro 1.7: La evaluación de costo-efectividad

Comparación de evaluaciones de programas que influyen en el aprendizaje en las escuelas primarias

Al evaluar un cierto número de programas con objetivos similares, es posible comparar la costo-efectividad relativa de diferentes enfoques para mejorar los resultados, como el aprendizaje en las escuelas primarias. Para que esto sea posible, los evaluadores deben divulgar no solo los resultados de la evaluación de impacto sino también información detallada sobre el costo de las intervenciones. En un meta análisis de los resultados de aprendizaje efectuado en países en desarrollo, Kremer, Brannen y Glennerster (2013) utilizaron información sobre el costo de 30 evaluaciones de impacto para analizar la costo-efectividad de diferentes tipos de intervenciones educativas.

Los autores compararon varios tipos de intervenciones en educación, entre ellas el acceso a la educación, los insumos habituales, las innovaciones pedagógicas, la rendición de cuentas de los maestros y la gestión basada en la escuela. En particular, investigaron las mejoras en las puntuaciones de las pruebas, en términos de desviaciones estándar, que podían obtenerse por cada US$100 invertidos en el programa. Aunque es probable que los costos disminuyeran si los programas se implementaban a escala, para mayor consistencia, los investigadores utilizaron los costos tal como se informaba

Continúa en la página siguiente.

Recuadro 1.7: La evaluación de costo-efectividad *(continúa)*

en las evaluaciones. Así, encontraron que las reformas pedagógicas y las intervenciones que mejoran la rendición de cuentas y aumentan los incentivos para los maestros tienden a ser las más costo-efectivas. Por otro lado, llegaron a la conclusión de que proveer más de los mismos insumos sin cambiar la pedagogía o la rendición de cuentas tenía impactos limitados en las puntuaciones de las pruebas. Por ejemplo, un programa aplicado en Kenia que incrementó el número de maestros en las escuelas no tuvo impactos significativos en las puntuaciones de las pruebas de los alumnos.

Los programas que empoderaban a las comunidades locales a través de intervenciones de gestión basadas en la escuela parecían ser los más exitosos y costo-efectivos, sobre todo cuando estas reformas se formalizaron. Por ejemplo, si bien la creación y formación de comités de escuelas locales en Indonesia no tuvo impactos significativos en las puntuaciones de las pruebas, lograr que los comités fueran más representativos a través de las elecciones resultó sumamente costo-efectivo.

Como ilustra su estudio, comparar las evaluaciones de intervenciones que tienen objetivos similares puede arrojar luz sobre la efectividad de diferentes intervenciones en diferentes contextos. Sin embargo, los investigadores deben reconocer que los contextos varían de forma considerable según los programas y escenarios. También sigue siendo relativamente raro contar con abundancia de datos de distintos programas en términos de mediciones, evaluaciones de impacto e información del costo de resultados comparables.

Fuente: Kremer, Brannen y Glennerster (2013).

Consideraciones éticas con respecto a la evaluación de impacto

Cuando se toma la decisión de diseñar una evaluación de impacto, se deben considerar algunos asuntos éticos. Se han formulado preguntas a propósito de si la evaluación de impacto es ética en sí misma y por sí sola. Un punto de partida de este debate consiste en considerar la ética de invertir recursos públicos en programas cuya efectividad se desconoce. En este contexto, la falta de evaluación puede en sí misma ser no ética. La información sobre la efectividad del programa que generan las evaluaciones de impacto puede conducir a una inversión más efectiva y ética de los recursos públicos.

Otras consideraciones éticas tienen que ver con las reglas utilizadas para asignar los beneficios del programa, con los métodos con los que se estudia a los seres humanos y con la transparencia en la documentación de los planes de investigación, datos y resultados. Estos temas se abordarán en detalle en el capítulo 13.

El principio ético más básico en una evaluación es que la prestación de intervenciones con beneficios conocidos no debería negarse o retrasarse

únicamente en función de los objetivos de la evaluación. En este libro se sostiene que las evaluaciones no deberían dictar cómo se asignan los beneficios, sino más bien que deberían ajustarse a reglas de asignación del programa que sean equitativas y transparentes. En este contexto, cualquier preocupación ética a propósito de las reglas de asignación del programa no nace de la evaluación de impacto en sí misma sino directamente de las reglas de operación del programa. Planificar las evaluaciones puede ser útil para dilucidar las reglas de operación del programa, y contribuir a estudiar si son equitativas y transparentes, sobre la base de criterios claros de elegibilidad.

La asignación aleatoria de los beneficios del programa a menudo suscita inquietudes éticas a propósito de la negación de dichos beneficios a destinatarios elegibles. Sin embargo, la mayoría de los programas funciona en contextos operativos con recursos financieros y administrativos limitados, lo cual impide llegar a todos los beneficiarios elegibles de una sola vez. Desde una perspectiva ética, todos los sujetos que son igualmente elegibles para participar en cualquier tipo de programa social deberían tener la misma oportunidad de ser destinatarios del programa. La asignación aleatoria cumple este requisito ético. En situaciones en las cuales un programa se aplicará en fases a lo largo del tiempo, la implementación se puede basar en seleccionar aleatoriamente el orden en que los beneficiarios, todos igualmente meritorios, serán receptores del programa. En estos casos, los destinatarios que ingresen más tarde en el programa pueden conformar el grupo de comparación para los primeros beneficiarios, generando un sólido diseño de evaluación, así como un método transparente e imparcial para asignar los escasos recursos.

La ética de la evaluación de impacto excede a la ética de las reglas de asignación del programa. También incluye la ética de la investigación en seres humanos, así como la ética de llevar a cabo investigaciones transparentes, objetivas y reproducibles, como se analiza en el capítulo 13.

En numerosos países e instituciones internacionales, se han creado juntas de revisión institucional o comités éticos para regular las investigaciones que involucran a seres humanos. Estos organismos se encargan de asesorar, aprobar y monitorear los estudios de investigación, con los objetivos fundamentales de proteger los derechos y promover el bienestar de todos los sujetos. Aunque las evaluaciones de impacto son eminentemente empresas operativas, también constituyen estudios de investigación y, como tales, deberían adherir a las directrices de investigación para seres humanos.

Conseguir que una evaluación de impacto sea objetiva, transparente y reproducible es un componente ético igualmente importante de la investigación. Para que la investigación sea transparente, los planes de evaluación de impacto pueden incluirse en un plan de análisis previo y ser sometidos a un registro de estudios. Una vez que la investigación se lleve a cabo, los datos

y códigos utilizados en el análisis pueden hacerse públicamente disponibles de modo que otras personas puedan replicar el trabajo, a la vez que se protege el anonimato.

La evaluación de impacto en las decisiones de políticas

Las evaluaciones de impacto son necesarias para informar a los responsables de las políticas en relación con una gama de decisiones, que abarcan desde los recortes de programas ineficientes hasta el aumento de escala de intervenciones que funcionan, o ajustar los beneficios del programa y seleccionar entre diversas alternativas de programas. Dichas evaluaciones son más efectivas cuando se aplican de manera selectiva para responder a preguntas clave de políticas, y se suelen aplicar a programas piloto innovadores que están probando un enfoque desconocido pero prometedor. La evaluación de las transferencias condicionadas en México, que se describe en el recuadro 1.1, se volvió influyente no solo debido a la naturaleza innovadora del programa sino también porque la evaluación misma proporcionó evidencia creíble y sólida que no podía ignorarse en las posteriores decisiones de las políticas. La adopción y ampliación del programa tanto a nivel nacional como internacional tuvieron una fuerte influencia de los resultados de la evaluación.

Las evaluaciones de impacto se pueden utilizar para explorar diferentes tipos de preguntas relacionadas con las políticas. La forma básica de evaluación de impacto probará la efectividad de un determinado programa. En otras palabras, responderá a la pregunta: ¿son un determinado programa o una determinada intervención efectivos en comparación con la ausencia del programa? Como se verá en la parte 2 del libro, este tipo de evaluación de impacto depende de la comparación entre un grupo de tratamiento al que se aplicó la innovación, el programa o la política y un grupo al que no se le aplicó, con el fin de estimar la efectividad. El principal reto en una evaluación de impacto consiste en construir un grupo de comparación que sea lo más similar posible al grupo de tratamiento. El grado de comparabilidad entre los grupos de tratamiento y comparación es central para la "validez interna" de la evaluación y, por lo tanto, es fundamental para evaluar el impacto causal de un programa.

Las evaluaciones de impacto también se están utilizando cada vez más para probar innovaciones de diseño en un programa sin un grupo de comparación "puro" seleccionado fuera del programa. Estos tipos de evaluaciones a menudo se realizan para ver si una determinada innovación de diseño puede mejorar la efectividad del programa o disminuir los costos (véase el recuadro 1.8).

Recuadro 1.8: Evaluación de programas innovadores

El equipo de Beahavioural Insights del Reino Unido

Creado en 2010 por el gobierno británico, el equipo de Behavioural Insights (BIT, por sus siglas en inglés) fue la primera institución estatal destinada a mejorar los servicios públicos a través de la aplicación de la ciencia del comportamiento. Los objetivos de la organización son mejorar la costo-efectividad de los servicios públicos, introducir modelos realistas de comportamiento humano en los análisis de las políticas y permitir que las personas tomen mejores decisiones. Con este objetivo, el BIT utiliza experimentos con evaluaciones de impacto incorporadas para probar ideas innovadoras en las políticas públicas. Desde su creación, la organización ha implementado más de 150 pruebas de control aleatorizado en una amplia variedad de ámbitos de las políticas nacionales, a menudo utilizando datos administrativos.

El BIT ha llevado a cabo evaluaciones de innovaciones en los servicios públicos sobre la base de la literatura de la ciencia del comportamiento. La organización colaboró con un municipio de Londres para introducir un incentivo de sorteo para mejorar la inscripción de los votantes antes de las elecciones. Los residentes fueron asignados aleatoriamente a tres grupos: i) sin sorteo, ii) un sorteo con un premio de £1.000 si se inscribían hasta cierta fecha y iii) un sorteo con un premio de £5.000 si se inscribían antes de esa misma fecha. El BIT llegó a la conclusión de que el incentivo del sorteo aumentó de manera significativa la inscripción de los votantes. Además, ahorró al gobierno local mucho dinero; anteriormente, el gobierno había recurrido a una onerosa campaña puerta a puerta para incrementar la inscripción de votantes.

En otra evaluación innovadora, el BIT se asoció con el Servicio Nacional de Salud y el Departamento de Salud para analizar cómo animar en términos costo-efectivos a las personas a registrarse como donantes de órganos. Se trata de uno de los ensayos aleatorios controlados más grandes jamás llevados a cabo en el sector público del Reino Unido. Los investigadores encontraron resultados alentadores a partir de una intervención que probaba el uso de diferentes mensajes en una página web pública de alto tránsito. La frase breve con mejores resultados se basó en la idea de reciprocidad y preguntaba: "Si necesitara un trasplante de órganos, ¿recurriría a él? Si la respuesta es sí, ayude a otros".

El BIT es de propiedad conjunta y está financiado por el gobierno británico, Nesta (una institución de beneficencia para la innovación), y los propios empleados. El modelo se ha ampliado fuera del Reino Unido y se han creado oficinas de BIT en Australia y Estados Unidos. Además, Estados Unidos siguió el modelo BIT para crear una iniciativa social y de ciencia del comportamiento en la Casa Blanca en 2015.

Fuente: Behavioural Insights Team.

Las evaluaciones también pueden utilizarse para probar la efectividad de las alternativas de implementación de un programa. Por ejemplo, pueden responder a la siguiente pregunta: cuando un programa se puede implementar de diversas maneras, ¿cuál es la modalidad de programa más efectiva o la más costo-efectiva? En este tipo de evaluación pueden compararse dos o más

enfoques o rasgos de diseño dentro de un programa para generar evidencia en relación con cuál es la alternativa más costo-efectiva para lograr un determinado objetivo. A estas alternativas de programa suele denominárselas "ramas de tratamiento". Por ejemplo, puede que un programa desee probar campañas de extensión alternativas y seleccione a un grupo para que reciba una campaña de correo, mientras que otro es destinatario de visitas puerta a puerta y un tercero recibe mensajes de texto SMS, para evaluar cuál es el método más costo-efectivo. Las evaluaciones de impacto que prueban tratamientos de programas alternativos suelen incluir un grupo de tratamiento para cada una de las ramas de tratamiento, así como un grupo de comparación "puro" que no recibe ninguna intervención del programa. Este tipo de evaluaciones permite que los responsables de la toma de decisiones elijan entre distintas alternativas de implementación, y puede ser muy útil para mejorar el desempeño de los programas y ahorrar costos (recuadro 1.9).

Recuadro 1.9: La evaluación de alternativas de diseño de programas

La desnutrición y el desarrollo cognitivo en Colombia

A comienzos de los años setenta, la Estación de Investigación de Ecología Humana, en colaboración con el Ministerio de Educación de Colombia, implementó un programa piloto para abordar el problema de la desnutrición infantil en Cali, Colombia, mediante atención sanitaria y actividades educativas, además de alimentos y complementos nutricionales. Como parte del plan piloto, un equipo de evaluadores debía determinar cuánto duraría un programa de este tipo para reducir la desnutrición entre los niños en edad preescolar de las familias de bajos ingresos, y si las intervenciones también podrían generar mejoras en el desarrollo cognitivo.

Finalmente, el programa se implementó para ocho familias elegibles, aunque durante el plan piloto los evaluadores pudieron comparar grupos similares de niños que recibían el tratamiento durante lapsos diferentes. Primero, los evaluadores utilizaron un proceso de selección para identificar un grupo de 333 niños desnutridos. Estos niños fueron clasificados en 20 sectores por barrio, y cada sector fue asignado aleatoriamente a uno de cuatro grupos de tratamiento. Los grupos diferían solo en la secuencia en que comenzaban el tratamiento y, por lo tanto, en la cantidad de tiempo que dedicarían al programa. El grupo 4 fue el que empezó primero y se vio expuesto al tratamiento durante el período más largo, seguido de los grupos 3, 2 y 1. El tratamiento consistía en seis horas diarias de atención sanitaria y actividades educativas, más alimentos y complementos nutricionales. A intervalos regulares a lo largo del programa, los evaluadores utilizaron pruebas cognitivas para seguir el progreso de los niños en los cuatro grupos.

Los evaluadores llegaron a la conclusión de que los niños que estaban más tiempo

Continúa en la página siguiente.

Recuadro 1.9: La evaluación de alternativas de diseño de programas *(continúa)*

en el programa mostraban la mayor mejora en el área cognitiva. En el test de inteligencia Stanford-Binet, que calcula la edad mental menos la edad cronológica, el promedio de los niños del grupo 4 era de -5 meses y el de los niños del grupo 1 era de -15 meses.

Este ejemplo ilustra cómo los encargados de implementar el programa y los responsables de las políticas pueden utilizar las evaluaciones de múltiples ramas de tratamiento para determinar cuál es la alternativa más efectiva para un programa.

Fuente: McKay et al. (1978).

Además, se pueden hacer comparaciones entre subgrupos de receptores dentro de una determinada evaluación para responder a la siguiente pregunta: ¿el programa es más efectivo para un subgrupo que para otro subgrupo? Por ejemplo, la introducción de un nuevo programa de estudios, ¿aumentó más las puntuaciones de las pruebas entre las alumnas que entre los alumnos? Este tipo de preguntas de una evaluación de impacto se orienta a documentar si hay alguna heterogeneidad en los impactos del programa en diferentes subgrupos. Estas preguntas deben tenerse en cuenta al comienzo, dado que es necesario incorporarlas en el diseño de una evaluación de impacto y requieren muestras suficientemente grandes para llevar a cabo el análisis de los distintos subgrupos de interés.

Más allá de las diversas características del diseño ya tratadas, cabe considerar los canales a través de los cuales las evaluaciones de impacto influyen en las políticas públicas. Esto puede suceder en relación con decisiones acerca de continuar, reformar o poner fin a un programa. Los resultados de la evaluación de impacto también pueden fundamentar el aumento de la escala de las pruebas piloto, como queda ilustrado en el caso de Mozambique presentado en el recuadro 1.2.

Las evaluaciones también pueden aportar evidencia de un país a otro o se pueden utilizar para analizar cuestiones fundamentales, como las relacionadas con la conducta. Aventurarse más allá de las fronteras de una evaluación de un programa individual suscita la pregunta sobre su carácter generalizable. Como se verá en el capítulo 4, en el contexto de una determinada evaluación, la muestra de la evaluación está diseñada para ser estadísticamente representativa de la población de unidades elegibles de las que se extrae la propia muestra y, por lo tanto, es externamente válida. Pero más allá de la validez externa, el carácter generalizable determinará si los resultados de una evaluación realizada a nivel local serán válidos en otros entornos y para otros grupos de población. Este concepto más expansivo y ambicioso

depende de la acumulación de evidencia empírica creíble en toda una gama de entornos.

El campo de la evaluación de impacto se basa cada vez más en el creciente inventario de evaluaciones creíbles para alcanzar conclusiones ampliamente generalizables. Este esfuerzo se centra en probar si una teoría concreta del cambio es válida en diferentes contextos y si un programa similar probado en entornos diferentes arroja resultados similares (véase el recuadro 1.10). El uso de múltiples evaluaciones para responder a preguntas centrales o reunir evidencia a través de meta análisis, estudios sistemáticos y registros de evaluación está aumentando rápidamente, y abre una nueva frontera en el trabajo de evaluación. Si los resultados son consistentes en diferentes y múltiples entornos, esto brinda mayor confianza a los responsables de las políticas en cuanto a la viabilidad del programa en toda una gama de contextos y de grupos de población. Se trata de una consideración importante, dado que los debates acerca de la capacidad de replicar resultados son fundamentales en cuestiones relacionadas con la efectividad y escalabilidad más amplia de un determinado programa.

Recuadro 1.10: El enfoque de evaluaciones de impacto de *clusters*

Construcción estratégica de la evidencia para reducir las brechas de conocimiento

A pesar de que el carácter generalizable de una sola evaluación de impacto puede ser bajo, en combinación con evaluaciones similares en diferentes contextos los profesionales del desarrollo pueden elaborar conclusiones más ampliamente aplicables a propósito de qué funciona y qué no. Cada vez más, las iniciativas de evaluación de impacto como el Fondo Estratégico para la Evaluación de Impacto del Banco Mundial (SIEF) y la Evaluación de Impacto del Desarrollo (DIME), así como la Iniciativa Internacional para la Evaluación de Impacto (3IE), se proponen brindar a los responsables de las políticas perspectivas para aplicar más ampliamente las intervenciones de un programa y de las políticas, utilizando un enfoque de "*cluster* de investigación".

A menudo, las convocatorias de propuestas se orientan en torno a un conjunto de preguntas de investigación destinadas a fundamentar el programa y el diseño de las políticas, para generar evaluaciones de impacto que contribuirán a una base de evidencia coordinada. El objetivo consiste en orientar la investigación y la generación de evidencia en torno a tipos de intervenciones o tipos de resultados.

Dentro de estos *clusters* se producen evaluaciones para llenar lagunas en el conjunto

Continúa en la página siguiente.

Recuadro 1.10: El enfoque de evaluaciones de impacto de *clusters* *(continúa)*

de pruebas existente. Por ejemplo, hay sólida evidencia que demuestra que los niños que reciben una combinación de nutrición, estimulación cognitiva y apoyo sanitario en los primeros 1.000 días de vida tienen más probabilidades de evitar retrasos en el desarrollo. Sin embargo, faltan investigaciones sobre cuál es la mejor manera de prestar este apoyo combinado de formas escalables y costo-efectivas. SIEF apoya la investigación para explorar esta pregunta en Bangladesh, Colombia, India, Indonesia, Madagascar, Mozambique, Nepal y Níger.

Agrupar las evaluaciones en torno a un cúmulo común de preguntas de investigación y utilizando una batería clave de instrumentos para medir los resultados permite a los responsables de las políticas y a los profesionales del desarrollo ver qué tipos de programas funcionan en múltiples entornos. Después, podrán revisar sus propios diseños de políticas y programas con un sentido más afinado de los contextos en que determinados programas han funcionado o no, o teniendo en cuenta cómo en diversos casos se han logrado resultados concretos.

Fuentes: DIME (www.worldbank.org/dime); SIEF (http://www.worldbank.org/en/programs/sief-trust-fund); 3IE (http://www.3ieimpact.org).

La decisión de llevar a cabo una evaluación de impacto

No todos los programas justifican una evaluación de impacto. Las evaluaciones de impacto deberían utilizarse selectivamente cuando la pregunta que se plantea exige un exhaustivo análisis de la causalidad. Estas evaluaciones pueden ser costosas si uno tiene que recopilar sus propios datos, y el presupuesto con el que se cuenta para la evaluación debería utilizarse de manera estratégica. Si uno está comenzando, o pensando en ampliar un nuevo programa y tiene dudas acerca de proceder o no con una evaluación de impacto, formularse unas pocas preguntas básicas le ayudará en esta decisión.

La primera pregunta que debe formularse es: ¿qué está en juego? ¿Será que la evidencia del éxito del programa, o la modalidad del mismo o la innovación en el diseño fundamentarán decisiones clave? Estas decisiones a menudo implican asignaciones presupuestarias y determinan la escala del programa. Si el presupuesto es limitado o si los resultados afectarán solo a unas pocas personas, puede que una evaluación de impacto no merezca la pena. Por ejemplo, puede que no se justifique llevar a cabo una evaluación de impacto de un programa en una clínica pequeña que ofrece asesoría a los pacientes de hospital mediante voluntarios. En cambio, una reforma del salario de los maestros que eventualmente afectará a todos los maestros de

primaria del país sería un programa con elementos mucho más importantes en juego.

Si se decide que hay mucho en juego la siguiente pregunta es: ¿existe evidencia que demuestre que el programa funciona? Concretamente, ¿se sabe cuál sería el alcance del impacto del programa? ¿Hay evidencia disponible de programas similares en circunstancias similares? Si no hay evidencia disponible acerca del potencial del tipo de programa proyectado, puede que convenga comenzar con una prueba piloto que incorpore una evaluación de impacto. En cambio, si hay evidencia disponible de circunstancias similares, el costo de una evaluación de impacto probablemente estaría justificado solo si puede abordar una nueva pregunta determinante para las políticas públicas. Este sería el caso si el programa incluye innovaciones sustanciales que todavía no han sido probadas.

Para justificar la movilización de recursos técnicos y financieros necesarios para llevar a cabo una evaluación de impacto de alta calidad, la intervención que será evaluada debe ser:

- *Innovadora*. Probará un enfoque nuevo y prometedor.
- *Aplicable*. Se puede aumentar la escala o se puede aplicar en un entorno diferente.
- *Estratégicamente relevante*. La evidencia proporcionada por la evaluación de impacto fundamentará una decisión clave ligada a la intervención. Esto podría estar relacionado con la ampliación del programa, reformas o asignaciones presupuestarias.
- *No probada*. Se sabe poco acerca de la efectividad del programa o de las alternativas de diseño, tanto a nivel global como en un contexto específico.
- *Influyente*. Los resultados se utilizarán para fundamentar decisiones de políticas.

Una pregunta final es la siguiente: ¿se cuenta con los recursos necesarios para una buena evaluación de impacto? Estos recursos atañen a elementos técnicos, como datos y el tiempo adecuado, recursos financieros para llevar a cabo la evaluación y recursos institucionales de los equipos que participan, más su interés y compromiso para construir y utilizar evidencia causal. Como se aborda en profundidad en el capítulo 12, un equipo de evaluación es esencialmente una asociación entre dos grupos: un equipo de responsables de las políticas y un equipo de investigadores. Los equipos tienen que trabajar en aras del objetivo común de asegurar que una evaluación técnicamente robusta y bien diseñada se implemente de manera adecuada y arroje resultados relevantes para cuestiones clave de las políticas y del diseño del programa.

Una clara comprensión de la premisa y de la promesa de la evaluación de impacto por parte del equipo de evaluación contribuirá a asegurar su éxito.

Si usted decide que tiene sentido encarar una evaluación de impacto, en virtud de las preguntas planteadas y la necesidad relacionada de analizar la causalidad, más los elementos en juego asociados con los resultados y la necesidad de evidencia acerca del desempeño de su programa, entonces lo invitamos a continuar con la lectura. Este libro está dedicado a usted y a su equipo de evaluación.

Recursos adicionales

- Para material relacionado con este capítulo e hipervínculos de recursos adicionales, se recomienda consultar el sitio web de Evaluación de Impacto en la Práctica (www.worldbank.org/ieinpractice).
- Para más información sobre las evaluaciones de impacto, véase S. R. Khandker, G. B. Koolwal y H. A. Samad (2009), *Handbook on Quantitative Methods of Program Evaluation*. Washington, D.C.: Banco Mundial.
- Para un buen resumen de las pruebas controladas aleatorias, véase R. Glennerster y K. Takavarasha (2013), *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press.
- Otros recursos sobre pruebas controladas aleatorias:
  - ➢ E. Duflo, R. Glennerster y M. Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit." Documento de discusión CEPR Núm. 6059. Londres: Center for Economic Policy Research.
  - ➢ E. Duflo y M. Kremer (2008), "Use of Randomization in the Evaluation of Development Effectiveness." En: *Evaluating Development Effectiveness* (vol. 7). Washington, D.C.: Banco Mundial.
- Otros recursos útiles sobre evaluación de impacto:
  - ➢ F. Leeuw y J. Vaessen (2009), *Impact Evaluations and Development. NONIE Guidance on Impact Evaluation*. Washington, D.C.: NONIE.
  - ➢ M. Ravallion (2001), "The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation." *World Bank Economic Review* 15 (1): 115–40.
  - ➢ ——. 2008. "Evaluating Anti-Poverty Programs." En: *Handbook of Development Economics* (vol. 4), editado por Paul Schultz y John Strauss. Ámsterdam: North Holland.
  - ➢ ——. 2009. "Evaluation in the Practice of Development." *World Bank Research Observer* 24 (1): 29–53.

Notas

1. Para una visión general de los programas de transferencias condicionadas y el influyente rol que desempeña el programa de México, así como también su evaluación de impacto, véase Fiszbein y Schady (2009).

2. Los datos administrativos son aquellos datos recopilados rutinariamente como parte de la administración del programa e incluyen información sobre costos, registros y transacciones, normalmente como parte de la prestación de servicios.
3. Hay numerosas tipologías para evaluaciones y preguntas de las evaluaciones. Véanse Berk y Rossi (1998) y Rossi, Lipsey y Freeman (2003).
4. Los métodos "cuasi experimentales" son métodos de evaluación de impacto que utilizan un contrafactual, pero se diferencian de los métodos "experimentales" en el sentido de que no se basan en la asignación aleatoria de la intervención. Véase la sección 2 para un debate sobre ambos tipos de métodos.
5. Para una visión general de los métodos de investigación cualitativos, véase Patton (1990).
6. Adaptado del Bureau of Justice Assistance (1997: 97–98 y 102–03).
7. Para un debate detallado sobre el análisis de costo-beneficio, véanse Zerbe y Dively (1994); Brent (1996); Belli et al. (2001), y Boardman et al. (2001).

Referencias bibliográficas

Ananthpur, K., K. Malik y V. Rao. 2014. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." Documento de trabajo de investigación 6958. Washington, D.C.: Banco Mundial.

Bamberger, M., V. Rao y M. Woolcock. 2010. "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development." Documento de trabajo de investigación de políticas Núm. 5245. Washington, D.C.: Banco Mundial.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, et al. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348 (6236). doi:10.1126/science.1260799.

Behrman, J. R. y J. Hoddinott. 2001. "An Evaluation of the Impact of PROGRESA on Pre-school Child Height." FCND Briefs 104, International Food Policy Research Institute, Washington, D.C.

Belli, P., J. Anderson, H. Barnum, Jo. Dixon y J. P. Tan. 2001. *Handbook of Economic Analysis of Investment Operations*. Washington, D.C.: Banco Mundial.

Berk, R. A. y P. Rossi. 1998. *Thinking about Program Evaluation 2* (2da. edición). Thousand Oaks, CA: Sage Publications.

Berman, D. 2014. "Tanzania: Can Local Communities Successfully Run Cash Transfer Programs?" Washington, D.C.: Human Development Network, Banco Mundial.

Boardman, A., A. Vining, D. Greenberg y D. Weimer. 2001. *Cost-Benefit Analysis: Concepts and Practice*. New Jersey: Prentice Hall.

Bourguignon, F., F. H. G. Ferreira y P. G. Leite. 2003. "Conditional Cash Transfers, Schooling y Child Labor: Micro-Simulating Brazil's Bolsa Escola Program." *The World Bank Economic Review* 17 (2): 229–54.

BRAC (Bangladesh Rural Advancement Committee). 2013. "An End in Sight for Ultra-poverty." Nota de información de BRAC (noviembre). Disponible en http://www.brac.net/sites/default/files/BRAC%20Briefing%20-%20TUP.pdf.

Brent, R. 1996. *Applied Cost-Benefit Analysis.* Cheltenham, Reino Unido: Edward Elgar.

Bureau of Justice Assistance. 1997. *Urban Street Gang Enforcement*. Informe preparado por el Institute for Law and Justice, Inc. Washington, D.C.: Office of Justice Programs, Bureau of Justice Assistance, U.S. Department of Justice.

Creswell, J. W. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage Publications.

Evans, D. K., S. Hausladen, K. Kosec y N. Reese. 2014. "Community-based Conditional Cash Transfers in Tanzania: Results from a Randomized Trial." Washington, D.C.: Banco Mundial.

Fiszbein, A. y N. Schady. 2009. *Conditional Cash Transfers, Reducing Present and Future Poverty.* Documento de trabajo de investigación de políticas Núm. 47603. Washington, D.C.: Banco Mundial.

Gertler, P. J. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review* 94 (2): 336–41.

Glennerster, R. y K. Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton, NJ: Princeton University Press.

Imas, L. G. M. y R. C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, D.C.: Banco Mundial.

Kremer, M., C. Brannen y R. Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 (6130): 297–300.

Khandker, S., G. B. Koolwal y H. A. Samad. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices.* Washington, D.C.: Banco Mundial.

Levy, S. y E. Rodríguez. 2005. *Sin herencia de pobreza: el programa Progresa-Oportunidades de México.* Washington, D.C.: BID.

Martínez, S., S. Nadeau y V. Pereira, 2012. "The Promise of Preschool in Africa: A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique." Washington, D.C.: Banco Mundial y Save the Children.

McKay, H., A. McKay, L. Siniestra, H. Gómez y P. Lloreda. 1978. "Improving Cognitive Ability in Chronically Deprived Children." *Science* 200 (21): 270–78.

Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods* (2da. edición). Newbury Park, CA: Sage.

Rao, V. y M. Woolcock. 2003. "Integrating Qualitative and Quantitative Approaches in Program Evaluation." En: F. J. Bourguignon y L. Pereira da Silva, *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, pp. 165–90. Nueva York: Oxford University Press.

Rossi, P., M. W. Lipsey y H. Freeman. 2003. *Evaluation: A Systematic Approach* (7ma. edición) Thousand Oaks, CA: Sage Publications.

Schultz, P. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.

Skoufias, E. y B. McClafferty. 2001. "Is Progresa Working? Summary of the Results of an Evaluation by IFPRI." Washington, D.C.: International Food Policy Research Institute.

Todd, P. y K. Wolpin. 2006. "Using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico." *American Economic Review* 96 (5): 1384–1417.

Zerbe, R. y D. Dively. 1994. *Benefit Cost Analysis in Theory and Practice*. Nueva York: Harper Collins Publishing.

CAPÍTULO 2

La preparación de una evaluación

Pasos iniciales

Este capítulo reseña los pasos iniciales que es preciso ejecutar para configurar una evaluación. Estos pasos son: i) construir una teoría del cambio que describa cómo se supone que el proyecto logrará los objetivos previstos; ii) elaborar una cadena de resultados que sirva como instrumento útil para esbozar la teoría del cambio; iii) especificar las preguntas de la evaluación; y iv) seleccionar los indicadores para evaluar el desempeño.

Los cuatro pasos son necesarios y es preferible implementarlos al inicio, es decir, cuando comienza a diseñarse el proyecto de evaluación del programa o de las reformas. Esto requiere la participación de todas las partes interesadas, desde los responsables de las políticas hasta los implementadores del proyecto, con el fin de forjar una visión común de los objetivos y de cómo serán alcanzados. Esta participación permitirá crear un consenso sobre el enfoque de la evaluación y las principales preguntas a responder, y reforzará los vínculos entre la evaluación, la implementación del programa y el diseño de políticas públicas. La aplicación de estos pasos contribuye a la claridad y la especificidad, que son útiles tanto para elaborar una buena evaluación de impacto como para diseñar e implementar un programa efectivo. Cada uno de los pasos está claramente definido y está articulado en el modelo lógico incorporado en la cadena de resultados desde la precisión en la determinación de

los objetivos y las preguntas hasta la definición de las ideas integradas en la teoría del cambio, y los resultados esperados por la implementación del programa. Se requiere una especificación clara de los indicadores concretos que se utilizarán para medir el éxito del programa, no solo para asegurar que la evaluación esté enfocada sino también que el programa tenga objetivos bien definidos. Esto también proporciona una base firme para anticipar los efectos producidos. Estos parámetros son esenciales para definir los elementos técnicos de la evaluación, incluyendo el tamaño de la muestra requerida para la evaluación y los cálculos de la potencia, como se analiza en el capítulo 15.

En la mayoría de las evaluaciones de impacto será importante incluir una evaluación de costo-beneficio, o costo-efectividad, como se indica en el capítulo 1. Los formuladores de política deberán estar atentos para saber qué programas o reformas son efectivos pero también cuál es su costo. Se trata de un aspecto crucial para fundamentar decisiones acerca de si es viable aumentar la escala de un programa y si es posible replicarlo, dos consideraciones importantes en las decisiones de políticas públicas.

Construcción de una teoría del cambio

Una *teoría del cambio* es la descripción de cómo se supone que una intervención conseguirá los resultados deseados. En ese sentido, expone la lógica causal de cómo y por qué un proyecto, una modalidad de programa o un diseño de innovación lograrán los resultados previstos. Debido al enfoque causal de la investigación, una teoría del cambio es la base de cualquier evaluación de impacto. Su construcción es uno de los primeros requisitos para el diseño del proyecto, ya que contribuye a especificar las preguntas de la investigación.

Las teorías del cambio describen una secuencia de eventos que generan resultados: analizan las condiciones y los supuestos necesarios para que se produzca el cambio, explicitan la lógica causal inscrita en el programa y trazan el mapa de las intervenciones del programa a lo largo de las vías lógicas causales. Configurar una teoría del cambio en conjunto con las partes interesadas puede clarificar y mejorar el diseño del programa. Esto es especialmente importante en los programas que pretenden influir en las conductas, pues las teorías del cambio pueden ayudar a determinar los insumos y actividades de la intervención, qué productos se generan y cuáles son los resultados finales derivados de los cambios de comportamiento de los beneficiarios.

El mejor momento para desarrollar una teoría del cambio es al comienzo de la fase de diseño, cuando es posible reunir a las partes interesadas con el fin de definir una visión colectiva del programa, sus objetivos y la ruta para alcanzar esos objetivos. Así, las partes interesadas podrán implementar el programa a partir de un entendimiento común del mismo, de sus objetivos y de su funcionamiento.

Por otra parte, es útil que los diseñadores de programas revisen la literatura en busca de evidencia que describa experiencias y programas similares, y comprueben los contextos y los supuestos detrás de las vías causales de la teoría del cambio que configuran. Por ejemplo, en el caso del proyecto de reemplazo de suelos de tierra por suelos de cemento desarrollado en México (que se reseña en el recuadro 2.1), la literatura aporta

Recuadro 2.1: La articulación de una teoría del cambio: de los pisos de cemento a la felicidad en México

En la evaluación del proyecto Piso Firme, Cattaneo et al. analizaron el impacto de la mejora de las viviendas en la salud y el bienestar. Tanto el proyecto como la evaluación estuvieron motivados por una clara teoría del cambio.

El objetivo del proyecto Piso Firme consiste en aumentar la calidad de vida, sobre todo en lo referente a la salud, de los grupos vulnerables que viven en zonas densamente pobladas y de bajos ingresos en México. El programa se inició en el estado norteño de Coahuila sobre la base de una evaluación contextual llevada a cabo por el gobierno estadual.

La cadena de resultados del programa es clara. Se realizan visitas puerta por puerta en los barrios elegidos para ofrecer a los hogares la construcción de 50 m² de suelo de cemento. El gobierno compra y entrega el cemento y los hogares y los voluntarios comunitarios aportan la mano de obra. El producto es la construcción de un suelo de cemento que se puede completar aproximadamente en un día. Entre los resultados previstos por este programa se destacan una mayor limpieza, una mejora en la salud de los habitantes de la casa y un aumento de su felicidad.

La lógica de esta cadena de resultados es que los suelos de tierra son un foco de enfermedades parasitarias porque es difícil mantenerlos limpios. Los parásitos viven y se multiplican en las heces y pueden ser ingeridos por las personas cuando son introducidos en el interior de la vivienda por los seres humanos o los animales. La evidencia demuestra que los niños pequeños que habitan en casas con suelos de tierra tienen más probabilidades de sufrir infecciones intestinales provocadas por parásitos, que pueden causar diarrea y desnutrición y que a menudo perjudican el desarrollo cognitivo e incluso pueden llevar a la muerte. Los suelos de cemento interrumpen la transmisión de las infecciones de los parásitos. También controlan la temperatura de manera más eficiente y mejoran el aspecto de la vivienda.

Estos resultados previstos sirvieron de fundamento para las preguntas que Cattaneo et al. (2009) formularon en la evaluación. La hipótesis del equipo era que al reemplazar los suelos de tierra con suelos de cemento se reduciría la incidencia de la diarrea, la desnutrición y la deficiencia de micronutrientes. A su vez, las mejoras en la salud y nutrición deberían impactar positivamente en el desarrollo cognitivo de los niños pequeños. Los investigadores también anticiparon y comprobaron un mayor bienestar entre los adultos, medido por el aumento de la satisfacción en la población respecto de las condiciones de sus viviendas y el descenso de las tasas de depresión y de estrés percibidas.

Fuente: Cattaneo et al. (2009).

información valiosa sobre cómo se transmiten los parásitos y de qué manera la infestación provocada por estos organismos produce diarrea infantil.

Desarrollo de una cadena de resultados

Una cadena de resultados es una manera de describir una teoría del cambio. Otros enfoques incluyen modelos teóricos, modelos lógicos, marcos lógicos y modelos de resultados. Todos estos modelos integran los elementos básicos de una teoría del cambio, a saber: una cadena causal, una especificación de las condiciones e influencias externas y la determinación de los supuestos clave. En este libro se utilizará el modelo de cadena de resultados porque es el más sencillo y claro para describir la teoría del cambio en el contexto operativo de los programas de desarrollo.

Concepto clave

Una cadena de resultados establece la secuencia de insumos, actividades y productos que previsiblemente mejorarán los resultados y los resultados finales.

Una cadena de resultados establece la lógica causal desde el inicio del programa, empezando con los recursos disponibles, hasta el final, teniendo en cuenta los objetivos de largo plazo. Fija una definición lógica y plausible de cómo una secuencia de insumos, actividades y productos relacionados directamente con el proyecto interactúa con el comportamiento y define las vías para lograr los impactos (véase el gráfico 2.1). Una cadena de resultados básica esquematizará un mapa con los siguientes elementos:

- *Insumos*. Los recursos de que dispone el proyecto, que incluyen el personal y el presupuesto.
- *Actividades*. Las acciones emprendidas o el trabajo realizado para transformar los insumos en productos.
- *Productos*. Los bienes y servicios tangibles que producen las actividades del programa (controlados de forma directa por la agencia ejecutora).
- *Resultados*. Los resultados que previsiblemente se lograrán cuando la población se beneficie de los productos del proyecto. En general, estos resultados se observan entre el corto y el mediano plazo y no suelen estar controlados de forma directa por la agencia ejecutora.
- *Resultados finales*. Los resultados finales alcanzados señalan si los objetivos del proyecto se cumplieron o no. Normalmente, los resultados finales dependen de múltiples factores y se producen después de un período más largo.

Tanto la implementación como los resultados forman parte de la cadena de resultados. La *ejecución* tiene que ver con el trabajo efectuado en el

Gráfico 2.1 Los elementos de una cadena de resultados

Fuente: Elaboración propia, sobre la base de múltiples fuentes.

proyecto, que incluye insumos, actividades y productos. Estos ámbitos, que son responsabilidad directa del proyecto, suelen ser monitoreados para verificar si el proyecto está generando los bienes y servicios previstos. Los *resultados* comprenden los resultados y los resultados finales, que no son controlados de manera directa por el proyecto y dependen de cambios en el comportamiento de los beneficiarios del programa, es decir, dependen de las interacciones entre la oferta (la implementación) y la demanda (los beneficiarios). Por lo general, estos ámbitos son objeto de evaluaciones de impacto para medir su efectividad.

Una buena cadena de resultados contribuirá a hacer aflorar los supuestos y riesgos implícitos en la teoría del cambio. Los formuladores de políticas están mejor situados para articular la lógica causal y los supuestos en los que descansa, así como los riesgos que pueden influir en el logro de los resultados previstos. El equipo que dirige la evaluación deberá explicitar los supuestos y riesgos implícitos en consulta con los responsables de las políticas. Una buena cadena de resultados también incluirá evidencia provista por la literatura relacionada con los resultados de programas similares.

Las cadenas de resultados son útiles para todos los proyectos, independientemente de que contemplen o no una evaluación de impacto, porque permiten a los responsables de las políticas y a los administradores del programa explicitar los objetivos del proyecto, lo que contribuye a clarificar la

lógica causal y la secuencia de eventos que se encuentran detrás de un programa. Además, pueden identificar brechas y eslabones débiles en el diseño del programa y, por lo tanto, pueden ayudar a mejorar su diseño. Las cadenas de resultados también facilitan el monitoreo y la evaluación pues especifican cuál es la información que debe ser monitoreada en cada eslabón de la cadena para realizar un seguimiento de la implementación del programa, y definen qué indicadores de resultados hay que incluir cuando se evalúa el proyecto.

La especificación de las preguntas de la evaluación

La claridad de la pregunta de la evaluación es el punto de partida de cualquier evaluación efectiva. La formulación de esta pregunta debe centrarse en la investigación, para asegurar que se ajusta al interés de las políticas en cuestión. En el caso de una evaluación de impacto, es necesario estructurarla como una hipótesis comprobable. La evaluación de impacto luego genera evidencia creíble para responder esa pregunta. Como se indicó antes, la pregunta fundamental de la evaluación de impacto es: ¿cuál es el impacto (o el efecto causal) de un programa en un resultado de interés? Se pone énfasis en el *impacto*, es decir, en los cambios *directamente atribuibles* a un programa, a una modalidad de programa o a una innovación de diseño.

La pregunta de la evaluación debe orientarse según el interés central de las políticas en cuestión. Como se señaló en el capítulo 1, las evaluaciones de impacto pueden analizar toda una gama de preguntas. En ese sentido, antes de estudiar cómo se llevará a cabo el proyecto, el equipo de evaluación debería aclarar cuál es la pregunta que se analizará como primer paso, sobre la base de la teoría del cambio.

Tradicionalmente, las evaluaciones de impacto se han centrado en el impacto que tiene un programa plenamente implementado en los resultados finales y en contraste con los resultados observados en un grupo de comparación que no ha sido beneficiado por el programa. Sin embargo, el uso actual de las evaluaciones de impacto se está ampliando. El equipo de evaluación puede inquirir: ¿la pregunta de evaluación clave es la pregunta "clásica" acerca de la efectividad de un programa para cambiar los resultados finales? ¿O se trata de probar si una modalidad de programa es más costo-efectiva que otra o bien de introducir una innovación en el diseño del programa que, de manera previsible, cambiará las conductas, como la matriculación? En la actualidad la evaluación de impacto está

incorporando nuevos enfoques, de manera creativa, para abordar las cuestiones de interés para el diseño de políticas, en una vasta gama de disciplinas (véase el recuadro 2.2).

En una evaluación de impacto, la pregunta de la evaluación debe ser formulada como una *hipótesis bien definida y comprobable*, que pueda cuantificar la diferencia entre los resultados obtenidos al contrastar los grupos de tratamiento y comparación. La cadena de resultados puede usarse como base para formular la hipótesis que se busca probar a partir de la evaluación de impacto. Como se señala en el recuadro 2.3, a menudo hay unas cuantas hipótesis asociadas con el programa, pero no es necesario explorar todas en una evaluación de impacto, y tampoco es posible hacerlo. En el ejemplo del programa de estudio de matemáticas que reseña el recuadro 2.2, la pregunta de la evaluación deriva de elementos fundamentales de la teoría del cambio y se formula como una hipótesis clara, comprobable y cuantificable: ¿cuál es el

Recuadro 2.2: Experimentos de mecanismo

Un *experimento de mecanismo* es una evaluación de impacto que prueba un mecanismo causal específico dentro de la teoría del cambio. Por ejemplo: se ha identificado un problema y se ha hallado un posible programa para remediarlo. Es preciso diseñar una evaluación a fin de probar la efectividad del programa. ¿La evaluación debería probar directamente el impacto del programa? Una corriente de pensamiento actual sostiene que una evaluación de programa de ese tipo quizá no sea siempre la mejor manera de comenzar y que, en algunos casos, puede ser preferible no llevar a cabo una evaluación de programa sino más bien poner a prueba algunos de los supuestos o mecanismos subyacentes. Los experimentos de mecanismo no prueban un programa; lo que prueban es un mecanismo causal que subyace a la elección de un programa.

Por ejemplo, un equipo de trabajo busca establecer si las personas que viven en barrios pobres de una ciudad tienen tasas de obesidad más altas que las personas que viven en sectores más acomodados de la misma ciudad. Después de realizar una investigación, el equipo observa que los barrios pobres cuentan con menos tiendas y puestos de frutas y verduras frescas y otros alimentos nutritivos y estima que esta falta de oferta puede estar contribuyendo a la obesidad, y que la situación se podría remediar mediante la entrega de subsidios a los fruteros a fin de que operen más puntos de venta. Una simple cadena de resultados podría tener el aspecto que se presenta en el gráfico B2.2.1.

Una evaluación del programa se centraría en probar el impacto de los subsidios a las fruterías en un conjunto de barrios pobres.

Continúa en la página siguiente.

Recuadro 2.2: Experimentos de mecanismo *(continúa)*

Gráfico B2.2.1 Identificación de un experimento de mecanismo en una cadena de resultados más larga

Fuente: Elaboración propia, sobre la base de múltiples fuentes.

En cambio, un experimento de mecanismo se enfocaría en evidenciar de manera más directa los supuestos subyacentes. Por ejemplo, puede poner a prueba el siguiente supuesto: si los habitantes de los barrios pobres tienen más acceso a alimentos nutritivos, comerán más de estos alimentos. Una forma de demostrarlo podría ser distribuir una canasta de frutas y verduras gratis una vez a la semana a un grupo de habitantes y comparar su consumo de frutas y verduras con el de los residentes que no reciben la canasta gratis. Si no se encuentran diferencias en el consumo de frutas y verduras en este experimento de mecanismo es probable que tampoco los subsidios a las fruterías tengan un impacto significativo debido a que uno de los mecanismos causales subyacentes no está funcionando.

En general, un experimento de mecanismo debería ser mucho más barato de implementar que una evaluación de programa completa, porque se puede llevar a cabo en una escala más pequeña. En el ejemplo anterior de la obesidad, proporcionar subsidios a los fruteros en numerosos barrios y supervisar a un gran número de residentes en esos barrios sería bastante caro, mientras que la entrega de la cesta con productos gratis resultaría mucho menos costosa y sería suficiente para contar con la participación de varios cientos de familias. Si el experimento de mecanismo demuestra que el mecanismo funciona, todavía habría que realizar un experimento de las políticas para evaluar si los subsidios son una manera efectiva de proveer frutas y verduras a los habitantes de los barrios pobres.

Fuente: Ludwig, Kling y Mullainathan (2011).

efecto del nuevo programa de matemáticas en las puntuaciones de las pruebas? En el ejemplo que se analiza a lo largo del libro, el Programa de Subsidios de Seguros de Salud, la pregunta de la evaluación es: ¿cuál es el efecto del Programa de Subsidios de Seguros de Salud en los gastos directos en salud de los hogares pobres?

Recuadro 2.3: Una reforma de las matemáticas en la enseñanza secundaria: elaboración de una cadena de resultados y una pregunta de la evaluación

El ministerio de Educación de un país X está pensando en introducir un nuevo programa de estudio de matemáticas en la enseñanza secundaria. La currícula ha sido diseñada para que resulte ser más accesible a los profesores y a los alumnos, a fin de mejorar el desempeño de los estudiantes en pruebas estandarizadas de matemáticas y, eventualmente, optimizar su capacidad para completar la escuela secundaria y tener acceso a mejores empleos. Esta cadena de resultados esboza la teoría del cambio del programa.

- Los insumos comprenden el personal del ministerio de Educación para liderar la reforma, los profesores de matemáticas de las escuelas secundarias, un presupuesto para desarrollar el nuevo programa de estudio, y las instalaciones municipales donde se impartirá la formación de los profesores de matemáticas.
- Las actividades del programa consisten en diseñar la nueva currícula de estudio de matemáticas, desarrollar un programa de capacitación de los profesores, capacitar a los profesores, y encargar la impresión y la distribución de los nuevos libros de texto.
- Los productos son el número de profesores que recibieron capacitación, el número de libros de texto entregados en las aulas y la adaptación de pruebas estandarizadas al nuevo programa de estudio.
- Los resultados en el corto plazo son la aplicación por parte de los profesores de los nuevos métodos, el uso de los libros de texto en las aulas y la administración de las nuevas pruebas.

Gráfico B2.3.1 Cadena de resultados para la reforma de la currícula de matemática en la escuela secundaria

Fuente: Elaboración propia, sobre la base de múltiples fuentes.

Continúa en la página siguiente.

Recuadro 2.3: Una reforma de las matemáticas en la enseñanza secundaria: elaboración de una cadena de resultados y una pregunta de la evaluación *(continúa)*

- Los resultados en el mediano plazo son las mejoras en el desempeño de los alumnos en las pruebas estandarizadas de matemáticas.
- Los resultados finales son el aumento en las tasas de finalización de los estudios secundarios y en las tasas de empleo, y el incremento en los ingresos de los graduados.

Diversas hipótesis sustentan la teoría del cambio:

- Los profesores que recibieron formación utilizan el nuevo programa de estudio de manera efectiva.
- Si los profesores reciben capacitación y se distribuyen los libros de texto, estos se emplearán y los alumnos seguirán el programa de estudio.
- El nuevo programa es superior al antiguo en la manera de impartir los conocimientos de matemáticas.
- Si la implementación se lleva a cabo como estaba previsto, los resultados de las pruebas de matemáticas mejorarán en un promedio de 5 puntos.
- El desempeño en las matemáticas en la secundaria influye en las tasas de finalización de la enseñanza secundaria, en las perspectivas de empleo y en el nivel de los ingresos.

La pregunta más importante de la evaluación, que ha sido desarrollada por el equipo de responsables de las políticas del Ministerio de Educación y los investigadores que participaron para determinar la efectividad del programa, es: ¿cuál es el efecto del nuevo programa de estudio de matemáticas en las puntuaciones de las pruebas? Esta pregunta apunta al corazón del interés de las políticas en cuanto a la efectividad del nuevo programa de estudios.

El Programa de Subsidios de Seguros de Salud: una introducción

El Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés, *Health Insurance Subsidy Program*) es un caso ficticio de un gobierno que emprende una reforma en gran escala del sector de la salud. Las preguntas relacionadas con este caso se utilizarán en todo el libro. El sitio web de Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice) contiene soluciones para las preguntas del estudio de caso del HISP, una base de datos y el código de análisis en Stata™, así como un manual técnico en línea que proporciona un tratamiento más formal del análisis de datos.

El objetivo final del HISP es mejorar la salud de la población del país. El innovador –y potencialmente caro– HISP se encuentra en etapa de pruebas. El gobierno está preocupado porque los hogares rurales pobres no pueden permitirse los costos de una atención sanitaria básica, lo cual

tiene consecuencias perjudiciales para su salud. A fin de abordar este problema, el HISP subsidia seguros de salud para los hogares rurales pobres, y cubre los costos relacionados con la atención primaria y los medicamentos. El propósito central del HISP consiste en reducir el costo de la atención sanitaria de las familias pobres y, eventualmente, mejorar los resultados de salud. Los responsables de las políticas están pensando en ampliar el HISP para cubrir al conjunto del país, lo cual costaría cientos de millones de dólares.

La cadena de resultados del HISP se ilustra en el gráfico 2.2. Las hipótesis relacionadas con la reforma del HISP son: i) los hogares se inscribirán en el programa una vez que se les ofrezca, ii) la inscripción en el programa disminuirá los gastos directos de los hogares en salud, iii) los costos impiden a la población rural tener acceso a la atención sanitaria y a los medicamentos disponibles, y iv) los gastos directos en los costos vinculados con la salud son un factor fundamental que contribuye a la pobreza y a los malos resultados de salud.

La pregunta clave de la evaluación es: ¿cuál es el impacto del Programa de Subsidios de Seguros de Salud en los gastos directos de los hogares en atención sanitaria? Tanto en el libro como en el material en línea, esta pregunta de la evaluación con relación al HISP será respondida varias veces, a partir de diferentes enfoques metodológicos. En ese marco, surgirán respuestas distintas –y a veces opuestas–, de acuerdo con la metodología de evaluación utilizada.

Gráfico 2.2 La cadena de resultados del HISP

La selección de indicadores de resultados y desempeño

Una pregunta clara de la evaluación debe ir acompañada de la especificación acerca de qué medidas de resultado se utilizarán para evaluar el desempeño, incluso en el caso de múltiples resultados. Las medidas de resultado seleccionadas se usarán para establecer si un programa o una reforma determinados tienen éxito o no. También son los indicadores que se pueden emplear como referencia al realizar los cálculos de la potencia con el fin de fijar los tamaños de la muestra necesarios para la evaluación, como se verá en el capítulo 15.

Luego de seleccionar los principales indicadores de interés, es preciso definir objetivos claros en lo relativo al éxito del programa. Este paso equivale a determinar el efecto anticipado del programa sobre los principales indicadores de resultado que se han seleccionado. Los *tamaños del efecto* son los cambios previstos como resultado del programa o de la reforma; por ejemplo, el cambio en las puntuaciones de las pruebas o en la tasa de adopción de un nuevo tipo de póliza de seguro. Los tamaños del efecto previstos son la base para llevar a cabo cálculos de la potencia (que se abordan con más detalles técnicos en el capítulo 15).

Es fundamental que los principales agentes interesados en el equipo de evaluación (tanto el equipo de investigación como el equipo de políticas públicas) estén de acuerdo tanto en los indicadores primarios de resultados de interés en la evaluación de impacto, como en los tamaños esperados de los efectos previstos como resultado del programa o de la innovación (para más detalles sobre el equipo de evaluación, véase el capítulo 12), ya que se usarán para juzgar el éxito del programa y formar la base de los cálculos de la potencia. Numerosas evaluaciones de impacto fracasan porque los tamaños de muestra no son lo bastante grandes para detectar los cambios generados por el programa: tienen un "déficit de potencia". En ese sentido, es sustancial definir los tamaños mínimos previstos del efecto para establecer criterios básicos del éxito del programa o de la innovación. Cuando hay datos disponibles, es posible realizar simulaciones ex ante para observar diferentes escenarios de resultados con el fin de establecer una referencia del tipo de tamaños del efecto que se pueden esperar en una gama de indicadores. Las simulaciones ex ante también se pueden utilizar para revisar indicadores iniciales de costo-beneficio o costo-efectividad y comparar intervenciones alternativas para provocar cambios en los resultados de interés.

Una cadena de resultados articulada con claridad proporciona un mapa útil para seleccionar los indicadores que se medirán a lo largo de la cadena,

a fin de monitorear la implementación del programa y evaluar los resultados. Como se ha dicho, es útil contar con la participación de agentes interesados del programa, tanto de los equipos de políticas como de los de investigación, para seleccionar estos indicadores y asegurar que son buenas medidas del desempeño del programa. La regla general indica que los criterios para asegurar que los indicadores utilizados son buenas medidas se resumen en el acrónimo EMARF. Los indicadores deberían ser:

Concepto clave
Los buenos indicadores son EMARF (específicos, medibles, atribuibles, realistas y focalizados).

- *Específicos*: para medir la información requerida de la forma más rigurosa posible.
- *Medibles*: para garantizar que la información se puede obtener fácilmente.
- *Atribuibles:* para asegurar que cada medida está relacionada con los logros del proyecto.
- *Realistas:* para garantizar que los datos se pueden obtener de manera oportuna, con una frecuencia y un costo razonables.
- *Focalizados:* en la población objetivo.

Al elegir los indicadores, es importante identificarlos a lo largo de toda la cadena de resultados, y no solo en el nivel de los resultados, de modo que puedan seguir la lógica causal de cualquier resultado observado del programa. En las evaluaciones de implementación que se centran en probar dos o más alternativas de diseño, los resultados de interés pueden producirse antes en la cadena de resultados, como un resultado adelantado o como resultado de una fase temprana. Aun cuando el interés solo esté puesto en las medidas de resultados para la evaluación, es sustancial realizar un seguimiento de los indicadores de implementación, de tal manera que se pueda determinar si las intervenciones se han llevado a cabo como estaban proyectadas, si han sido recibidas por los beneficiarios previstos y si han llegado a tiempo. Si no se identifican estos indicadores en toda la cadena de resultados se corre el riesgo de que la evaluación de impacto sea como una "caja negra" que podrá determinar si los resultados previstos se materializaron o no, pero no será capaz de explicar por qué.

Lista de verificación: datos para los indicadores

Como lista de verificación final, una vez que se han seleccionado los indicadores es útil pensar en las disposiciones para producir los datos con el fin de medir los indicadores. En el capítulo 4 se presenta un debate exhaustivo

sobre dónde conseguir los datos para la evaluación. Esta lista de verificación (adaptada de PNUD, 2009) abarca las disposiciones prácticas necesarias para asegurar que es posible producir todos los indicadores de manera fiable y oportuna:

- ✓ ¿Se han especificado con claridad los indicadores (productos y resultados)? Estos provienen de las preguntas clave de la evaluación y deberían ser consistentes con los documentos de diseño del programa y con la cadena de resultados.
- ✓ ¿Los indicadores son EMARF? Específicos, medibles, atribuibles, realistas y focalizados.
- ✓ ¿Cuál es la fuente de los datos de cada indicador? Es necesario definir con claridad la fuente de los datos, como una encuesta, un estudio o una reunión de las partes interesadas.
- ✓ ¿Con qué frecuencia se recopilarán los datos? Es preciso incluir un calendario.
- ✓ ¿Quién es el responsable de recopilar los datos? Se debe especificar quién es responsable de organizar la recopilación de datos, verificar la calidad y la fuente de los datos y asegurar el cumplimiento de las normas éticas.
- ✓ ¿Quién es responsable del análisis y de los informes? Hay que fijar la frecuencia de los análisis, el método de análisis y el responsable de los informes.
- ✓ ¿Qué recursos se necesitan para producir los datos? Es fundamental que los recursos requeridos sean claros y que estén destinados a producir los datos, que a menudo es la parte más cara de una evaluación si se recopilan datos primarios.
- ✓ ¿La documentación es adecuada? Es útil diseñar planes para documentar los datos, incluir la utilización de un registro y asegurar el anonimato.
- ✓ ¿Qué riesgos implica? Al realizar el monitoreo planificado y las actividades de evaluación es preciso considerar los riesgos y los supuestos, así como la manera en que pueden influir en la puntualidad y la calidad de los datos y de los indicadores.

Recursos adicionales

- Para obtener material complementario de este capítulo y para conseguir hipervínculos a recursos adicionales se recomienda consultar el sitio de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).

- El Módulo 1 del World Bank's Impact Evaluation Toolkit (http://www.worldbank.org/health/impactevaluationtoolkit) ofrece un gráfico de la teoría del cambio, una plantilla de cadena de resultados y ejemplos de indicadores para financiamiento basado en resultados.
- L. Morra Imas y R. Rist (2009) brindan un buen estudio de las teorías del cambio en "The Road to Results: Designing and Conducting Effective Development Evaluations." Washington, D.C.: Banco Mundial.
- Para un debate sobre cómo seleccionar los indicadores de desempeño, véanse:
  - ➢ L. Morra Imas y R. Rist (2009), "The Road to Results: Designing and Conducting Effective Development Evaluations." Washington, D.C.: Banco Mundial.
  - ➢ J. Kusek y R. Rist (2004), "Ten Steps to a Results-Based Monitoring and Evaluation System." Washington, D.C.: Banco Mundial.

Referencias bibliográficas

Cattaneo, M., S. Galiani, P. Gertler, S. Martinez y R. Titiunik. 2009. "Housing, Health and Happiness." *American Economic Journal: Economic Policy* 1 (1): 75–105.

Morra Imas, L. y R. Rist. 2009. "The Road to Results: Designing and Conducting Effective Development Evaluations." Washington, D.C.: Banco Mundial.

Kusek, J. y R. Rist. 2004. "Ten Steps to a Results-Based Monitoring and Evaluation System." Washington, D.C.: Banco Mundial.

Ludwig, J., J. Kling y S. Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.

PNUD (Programa de las Naciones Unidas para el Desarrollo). 2009. *Handbook on Planning, Monitoring and Evaluating for Development Results*. Nueva York: PNUD.

University of Wisconsin - Extension. 2010. "Enhancing Program Performance with Logic Models." Curso en línea. Disponible en http://www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html.

Vermeersch, C., E. Rothenbühler y J. Sturdy. 2012. "Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health." Washington, D.C.: Banco Mundial. Disponible en www.worldbank.org/health/impactevaluationtoolkit.

Segunda parte

CÓMO EVALUAR

La segunda parte de este libro explica cómo funcionan las evaluaciones de impacto, a qué preguntas responden, qué métodos están disponibles para llevarlas a cabo y cuáles son las ventajas y las desventajas de cada uno de ellos. El enfoque de la evaluación de impacto propuesto en este libro aboga por la selección del método más riguroso que sea compatible con las características operativas de un programa. El menú de opciones para una evaluación de impacto incluye la asignación aleatoria, las variables instrumentales, el diseño de regresión discontinua, las diferencias en diferencias y el pareamiento. Todos estos métodos comparten el objetivo común de construir grupos de comparación válidos que permitan estimar los verdaderos impactos de un programa.

El capítulo 3 introduce el concepto del *contrafactual* como piedra angular de la evaluación de impacto mediante una explicación de las propiedades que debe tener la estimación del contrafactual y ejemplos de estimaciones inválidas o falsas del contrafactual. Del capítulo 4 al 8 se aborda cada una de las metodologías para una evaluación de impacto: el capítulo 4 trata la asignación aleatoria; el 5, las variables instrumentales; el 6, el diseño de regresión discontinua;

el 7, las diferencias en diferencias, y el 8, el pareamiento. En estos capítulos se analiza cómo y por qué cada método puede producir una estimación válida del contrafactual, en qué contexto de las políticas públicas es posible implementarlos y cuáles son las principales limitaciones de cada uno. Asimismo, se ilustra el uso de los diferentes métodos con ejemplos específicos del mundo real de las evaluaciones de impacto que los han utilizado y con el estudio de caso del Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés, *Health Insurance Subsidy Program*) que se presentó en el capítulo 2. El capítulo 9 avanza sobre cómo abordar los problemas que pueden surgir durante la implementación, y reconoce que las evaluaciones de impacto no suelen ejecutarse siguiendo el diseño de un modo exacto. En este contexto, se consideran las dificultades habituales, entre ellas el cumplimiento imperfecto, los efectos secundarios y el desgaste de la muestra, y se ofrece orientación acerca de cómo afrontar estos problemas. El capítulo 10 concluye esta segunda parte del libro con una guía de las evaluaciones de programas multifacéticos, en especial aquellos con distintos niveles de tratamiento y múltiples brazos de tratamiento.

A lo largo de la segunda parte, el libro brinda la oportunidad de aplicar métodos y poner a prueba la comprensión a partir del estudio de caso del HISP. Como se recordará, la pregunta clave de la evaluación para los responsables de la política del HISP es: ¿qué impacto tiene este programa en los gastos directos en atención sanitaria de los hogares pobres? Se utilizará la base de datos del HISP para ilustrar cada método de evaluación e intentar responder esta pregunta. Siendo que ya se han reunido adecuadamente los datos, de modo que se han eliminado todos los problemas relacionados con estos datos, el libro proporcionará los resultados del análisis, que deberán ser interpretados. En ese sentido, la tarea del lector será determinar por qué la estimación del impacto del HISP cambia con cada método y decidir qué resultados son más fiables para justificar una decisión a favor o en contra de la ampliación del HISP. Las soluciones a las preguntas se hallan en el sitio web de evaluación de impacto en la práctica (www.worldbank.org/ieinpractice), donde, además, se encuentra la base de datos, el código de análisis en el *software* Stata™ y un manual técnico que proporciona un tratamiento más formal de los datos.

La parte 3 comienza indicando cómo usar las reglas de operación del programa, esto es: los recursos disponibles de un programa, los criterios para la selección de beneficiarios, y el plazo de implementación, como base para seleccionar un método de evaluación de impacto. Allí se presenta un marco sencillo para determinar cuál de las metodologías de evaluación de impacto expuestas en la parte 2 es más conveniente para un determinado programa, en función de sus normas operativas.

CAPÍTULO 3

Inferencia causal y contrafactuales

Inferencia causal

En las evaluaciones de impacto precisas y fiables existen dos conceptos esenciales: la inferencia causal y los contrafactuales.

Muchas preguntas de política económica tienen que ver con relaciones de causa y efecto: ¿la formación de los profesores mejora las puntuaciones de los alumnos en las pruebas? ¿Los programas de transferencias condicionadas consiguen mejores resultados en la salud de los niños? ¿Los programas de formación profesional aumentan los ingresos de quienes los han cursado?

Las evaluaciones de impacto pretenden responder estas preguntas de causa y efecto con precisión. Evaluar el impacto de un programa en un conjunto de resultados equivale a evaluar el efecto causal del programa en esos resultados.[1]

A pesar de que las preguntas de causa y efecto son habituales, contestarlas no es un asunto sencillo. En el contexto de un programa de formación profesional, por ejemplo, la sola observación de que los ingresos de una persona que ha recibido formación aumentan después de que ha completado ese programa no basta para establecer la causalidad. Tal vez los ingresos de esa persona se habrían incrementado aunque no hubiera sido objeto de la formación, sino como consecuencia de sus esfuerzos, de las condiciones

Concepto clave

Las evaluaciones de impacto establecen hasta qué punto un programa –*y solo ese programa*– provocó un cambio en un resultado.

cambiantes del mercado laboral o de muchos otros factores que influyen en los ingresos. Las evaluaciones de impacto ayudan a superar el problema de atribuir causalidad pues determinan, empíricamente, en qué medida un programa concreto –*y solo ese programa*– contribuye a cambiar un resultado. Para establecer causalidad entre un programa y un resultado se emplean métodos de evaluación de impacto a fin de descartar la posibilidad de que cualquier factor distinto del programa de interés explique el impacto observado.

La respuesta a la pregunta básica de la evaluación de impacto –cuál es el impacto o efecto causal de un programa (P) en un resultado de interés (Y)– se obtiene mediante la fórmula básica de la evaluación de impacto:

$$\Delta = (Y \mid P = 1) - (Y \mid P = 0)$$

Según esta fórmula, el impacto causal (Δ) de un programa (P) en un resultado (Y) es la diferencia entre el resultado (Y) con el programa (es decir, cuando $P = 1$) y el mismo resultado (Y) sin el programa (cuando $P = 0$).

Por ejemplo, si P representa un programa de formación profesional e Y simboliza los ingresos, el impacto causal de un programa de formación profesional (Δ) es la diferencia entre los ingresos de una persona (Y) después de participar en el programa de formación profesional (es decir, cuando $P = 1$) y los ingresos de la misma persona (Y) en el mismo momento en el tiempo, si no hubiera participado en el programa (cuando $P = 0$). Dicho de otro modo, se trata de medir el ingreso en el mismo momento en el tiempo para la misma unidad de observación (en este caso, una persona) pero en dos realidades diferentes. Si esto fuera posible, se observaría cuál sería el ingreso de ese mismo individuo en el mismo momento en el tiempo tanto con el programa como sin él, de modo que la única explicación posible de cualquier diferencia en los ingresos de esa persona sería el programa. Al comparar al mismo individuo consigo mismo en el mismo momento se conseguiría eliminar cualquier factor externo que también pudiera explicar la diferencia de los ingresos. En ese caso, sería posible confiar en que la relación entre el programa de formación profesional y el cambio en los ingresos es causal.

La fórmula básica de la evaluación de impacto es válida para cualquier unidad analizada, ya sea una persona, un hogar, una comunidad, una empresa, un colegio, un hospital u otra unidad de observación que pueda beneficiarse del programa o verse afectada por él. La fórmula también es válida para cualquier resultado (Y) relacionado con el programa en cuestión. Una vez que se han medido los dos componentes clave de esta fórmula –el resultado (Y) con el programa y sin él–, es posible responder cualquier pregunta acerca del impacto del programa.

El contrafactual

Como ya se señaló, es posible pensar en el impacto (Δ) de un programa como la diferencia en los resultados (Y) para la misma unidad (persona, hogar, comunidad, etc.) cuando ha participado en el programa y cuando no lo ha hecho. Sin embargo, es imposible medir al mismo sujeto en dos realidades diferentes al mismo tiempo. En cualquier momento del tiempo, un sujeto habrá participado en el programa o no lo habrá hecho. No se lo puede observar simultáneamente en dos realidades distintas (es decir, con el programa y sin él). Esto se denomina *problema contrafactual*: ¿cómo medir lo que habría ocurrido si hubieran prevalecido las otras circunstancias? Aunque se puede observar y medir el resultado (Y) para un participante del programa ($Y \mid P = 1$), no hay datos para establecer cuál habría sido su resultado en ausencia del programa ($Y \mid P = 0$). En la fórmula básica de la evaluación de impacto, *el término ($Y \mid P = 0$) representa el contrafactual*, lo cual se puede expresar como *¿cuál habría sido* el resultado si una persona no hubiera participado en el programa?

Concepto clave

El contrafactual es lo que habría ocurrido –cuál habría sido el resultado (Y) para un participante del programa– en ausencia del programa (P).

Por ejemplo, si el "señor Desafortunado" toma una píldora y muere cinco días después, el solo hecho de que el señor Desafortunado haya muerto después de tomar la píldora no permite concluir que la píldora haya sido la *causa* de su muerte. Quizá él estaba muy enfermo cuando tomó la píldora y fue la enfermedad la que provocó su muerte en lugar de la píldora. Para inferir la causalidad habrá que descartar todos los demás factores potenciales capaces de haber influido en el resultado en cuestión. En este sencillo ejemplo, para saber si la píldora causó la muerte del señor Desafortunado, un evaluador tendría que establecer qué le habría ocurrido al señor Desafortunado si no hubiera tomado la píldora. Como el señor Desafortunado tomó la píldora, no es posible observar de forma directa qué habría ocurrido si no lo hubiera hecho. Lo que le habría ocurrido si no hubiera tomado la píldora es el contrafactual. Para definir el impacto de la píldora, el principal reto del evaluador consiste en determinar qué aspecto tiene el estado contrafactual de la realidad para el señor Desafortunado (véase el recuadro 3.1).

Cuando se lleva a cabo una evaluación de impacto, es relativamente fácil obtener el primer término de la fórmula básica ($Y \mid P = 1$) –el resultado con un programa, también conocido como el resultado *bajo tratamiento*–, ya que basta con medir el resultado de interés para el participante del programa. Sin embargo, no es posible observar de forma directa el segundo término de la fórmula ($Y \mid P = 0$) para el participante. Es preciso obtener esta información *estimando el contrafactual*.

Concepto clave

Como no es posible observar directamente el contrafactual, es preciso estimarlo.

Para entender este concepto clave de estimación del contrafactual es útil recurrir a otro caso hipotético. La solución del problema contrafactual sería

Recuadro 3.1: El problema del contrafactual: la "señorita Única" y el programa de transferencias condicionadas

La "señorita Única" es una recién nacida cuya madre percibe una asignación monetaria mensual siempre que la niña sea sometida a chequeos regulares en el centro de salud local, reciba las vacunas y se chequee su crecimiento. Según el gobierno, la transferencia condicionada motivará a la madre de la señorita Única a acudir a los servicios de salud que requiere el programa y contribuirá al crecimiento normal y saludable de la niña. Para realizar una evaluación de impacto de la transferencia condicionada, el gobierno elige la altura como indicador de resultados de la salud en el largo plazo.

Idealmente, para evaluar el impacto del programa, habría que medir la altura de la señorita Única a los 3 años, cuando su madre recibió la transferencia condicionada y también cuando su madre no recibió dicha transferencia. Luego habría que comparar las dos alturas para establecer el impacto. Si fuera posible comparar la altura de la señorita Única a los 3 años bajo el programa con la altura de la señorita Única a los 3 años sin el programa se sabría que cualquier diferencia en la altura habría sido efecto solo del programa de transferencias condicionadas. Como todo lo demás relativo a la señorita Única sería igual, no habría otras características que explicaran la diferencia de altura.

Sin embargo, es imposible observar a la señorita Única con el programa de transferencias condicionadas y sin él: su familia cumple los requisitos (chequeos, vacunación, seguimiento del crecimiento) y recibe la transferencia condicionada o no lo hace. En otras palabras, no hay forma de observar cuál es el contrafactual. Como la madre de la señorita Única cumplió los requisitos y recibió la transferencia condicionada, no es factible saber qué altura tendría la señorita Única si su madre no hubiera recibido la transferencia condicionada.

Será difícil encontrar una comparación adecuada para la señorita Única porque, como su nombre indica, es única. Sus antecedentes socioeconómicos, sus atributos genéticos y sus características personales y del hogar no pueden ser hallados en ninguna otra persona. Si se compara la señorita Única con alguien que no participó en el programa –por ejemplo, el señor Inimitable–, la comparación tal vez resulte inadecuada: la señorita Única puede ser exactamente idéntica al señor Inimitable. Quizá la señorita Única y el señor Inimitable no tienen el mismo aspecto, no viven en el mismo lugar, no tienen los mismos padres y no midieron lo mismo cuando nacieron. Por lo tanto, si se observa que el señor Inimitable es más bajo que la señorita Única a los 3 años no será posible saber si la diferencia se debe al programa de transferencias condicionadas o a alguna de las muchas otras diferencias entre los dos niños.

posible si el evaluador pudiera encontrar un "clon perfecto" de un participante en el programa (véase el gráfico 3.1). Por ejemplo, si el señor Fulanito comienza a recibir US$12 como dinero de bolsillo y lo que se busca es medir el impacto de este tratamiento en su consumo de caramelos, la existencia de un clon perfecto del señor Fulanito haría la evaluación muy fácil, pues se

Gráfico 3.1 El clon perfecto

podría comparar el número de caramelos que come el señor Fulanito (por ejemplo, seis) cuando recibe el dinero, con el número de caramelos (por ejemplo, cuatro) que come su clon, que no recibe dinero. En este caso, el impacto del dinero de bolsillo sería de dos caramelos, es decir, la diferencia entre el número de caramelos consumidos bajo tratamiento (seis) y el número de caramelos consumidos sin tratamiento (cuatro). En realidad, está claro que es imposible hallar clones perfectos, y que incluso entre gemelos genéticamente idénticos hay diferencias importantes.

La estimación del contrafactual

La clave para estimar el contrafactual para los participantes del programa consiste en desplazarse del nivel individual o de la persona al nivel del grupo. A pesar de que no existe un clon perfecto de una persona única, es posible contar con propiedades estadísticas para generar dos grupos de personas que, si su número es lo bastante alto, sean indistinguibles una de otra desde el punto de vista estadístico en el nivel del grupo. El grupo que participa en el programa se denomina *grupo de tratamiento*, y su resultado es $(Y \mid P = 1)$ después de que ha participado en el programa. El *grupo de comparación* estadísticamente idéntico (a veces llamado "grupo de control") es el grupo que no es objeto del programa y permite estimar el resultado contrafactual $(Y \mid P = 0)$: es decir, el resultado que se habría obtenido en el grupo de tratamiento si no hubiera recibido el programa.

Concepto clave

Sin un grupo de comparación que produzca una estimación precisa del contrafactual, no se puede establecer el verdadero impacto de un programa.

Por lo tanto, en la práctica el reto de una evaluación de impacto es definir un grupo de tratamiento y un grupo de comparación que sean estadísticamente idénticos, en promedio, en ausencia del programa. Si los dos grupos son idénticos (estadísticamente), con la única excepción de que un grupo participa en el programa y el otro no, es posible estar seguros de que cualquier diferencia en los resultados tendría que deberse al programa. Encontrar esos grupos de comparación es la piedra angular de cualquier evaluación de impacto, al margen del tipo de programa que se evalúe. En pocas palabras, sin un grupo de comparación que produzca una estimación precisa del contrafactual, no se puede establecer el verdadero impacto de un programa.

En ese sentido, el principal desafío para identificar los impactos es crear un grupo de comparación válido que tenga las mismas características que el grupo de tratamiento en ausencia del programa. Concretamente, los grupos de tratamiento y de comparación deben ser iguales en al menos tres aspectos.

En primer lugar, las características promedio del grupo de tratamiento y del grupo de comparación deben ser idénticas en ausencia del programa.[2] Aunque no es necesario que las unidades individuales en el grupo de tratamiento tengan clones perfectos en el grupo de comparación, en promedio las características de los grupos de tratamiento y de comparación deberían ser las mismas. Por ejemplo, la edad promedio de las personas en el grupo de tratamiento debería ser la misma que en el grupo de comparación.

En segundo lugar, el tratamiento no tendría que afectar al grupo de comparación de forma directa ni indirecta. En el caso del señor Fulanito y el dinero de bolsillo, el grupo de tratamiento no debería transferir recursos al grupo de comparación (efecto directo) ni influir en el precio de los caramelos en los mercados locales (efecto indirecto). Por ejemplo, si lo que se busca es aislar el impacto del dinero de bolsillo en el consumo de caramelos, al grupo de tratamiento no se le deberían ofrecer más visitas a la tienda de caramelos que al grupo de comparación; de otra manera, no se podría distinguir si el consumo adicional de caramelos es consecuencia del dinero de bolsillo o del mayor número de visitas a la tienda de caramelos.

Concepto clave

Un grupo de comparación válido (1) tiene las mismas características, en promedio, que el grupo de tratamiento en ausencia del programa; (2) no es afectado por el programa; y (3) reaccionaría al programa de la misma manera que el grupo de tratamiento, si fuera objeto del programa.

En tercer lugar, los resultados de las personas en el grupo de control deberían cambiar de la misma manera que los resultados en el grupo de tratamiento, si ambos grupos son objeto del tratamiento (o no). En este sentido, los grupos de tratamiento y de comparación tendrían que reaccionar al programa de igual modo. Por ejemplo, si los ingresos de las personas del grupo de tratamiento aumentaran en US$100 gracias al programa de formación, los ingresos de las personas en el grupo de comparación también tendrían que subir US$100 si hubieran sido objeto de la formación.

Si se cumplen estas tres condiciones, solamente la existencia del programa de interés explicará cualquier diferencia en el resultado (Y) entre los dos grupos. Esto obedece a que la única diferencia entre los grupos de tratamiento y los de comparación es que los miembros del grupo de tratamiento recibieron el programa, mientras que los del grupo de comparación no lo recibieron. Cuando la diferencia en el resultado se puede atribuir totalmente al programa, se ha identificado el impacto causal del programa.

En el caso del señor Fulanito se observó que estimar el impacto del dinero de bolsillo en su consumo de caramelos exigía la tarea impracticable de encontrar el clon perfecto del señor Fulanito. En vez de analizar el impacto en un solo individuo, es más realista evaluar el impacto promedio en un grupo de individuos (véase el gráfico 3.2). En ese sentido, si se pudiera identificar otro grupo de individuos que comparten el mismo promedio de edad, composición por sexo, educación, preferencia por los caramelos, etc., con la salvedad de que no recibe el dinero de bolsillo adicional, sería posible estimar el impacto del dinero de bolsillo, pues este estaría conformado por la diferencia entre el consumo promedio de caramelos en ambos grupos. Por lo tanto, si el *grupo de tratamiento* consume una media de seis caramelos por persona, mientras que el *grupo de comparación* consume una media de cuatro, el impacto promedio del dinero de bolsillo adicional en el consumo de caramelos sería de dos caramelos.

Después de definir un *grupo de comparación válido*, es importante considerar qué ocurriría si la decisión fuera seguir adelante con una evaluación sin haber hallado ese grupo. Es evidente que un grupo de comparación no

Concepto clave

Cuando el grupo de comparación no estima con precisión el verdadero contrafactual, el impacto estimado del programa no es válido. En términos estadísticos es "sesgado".

Gráfico 3.2 Un grupo de comparación válido

válido difiere del grupo de tratamiento de alguna manera distinta de la ausencia de tratamiento. Debido a esas otras diferencias, la estimación de impacto puede ser no válida o, en términos estadísticos, puede ser *sesgada*: en ese caso la evaluación de impacto no estimará el verdadero impacto del programa, sino el efecto del programa mezclado con esas otras diferencias existentes entre los grupos.

Dos estimaciones falsas del contrafactual

En lo que queda de la segunda parte de este libro se abordarán los diversos métodos disponibles para construir grupos de comparación válidos que permitirán estimar el contrafactual. Sin embargo, antes resulta útil analizar dos métodos de uso habitual, aunque sumamente riesgosos, para construir grupos de comparación que a menudo conducen a estimaciones inadecuadas ("falsas") del contrafactual:

- *Las comparaciones antes-después* (también conocidas como comparaciones *pre-post* o *reflexivas*): cotejan los resultados del mismo grupo antes y después de participar en un programa.
- *Las comparaciones de inscritos y no inscritos (o autoseleccionados)*: comparan los resultados de un grupo que elige participar en un programa con otros de un grupo que elige no participar.

Contrafactual falso 1: comparación entre resultados antes y después de un programa

Una comparación antes-después intenta establecer el impacto de un programa a partir de un seguimiento de los cambios en los resultados en los participantes del programa a lo largo del tiempo. De acuerdo con la fórmula básica de la evaluación de impacto, el resultado para el grupo de tratamiento ($Y \mid P = 1$) es, sin duda, el resultado después de participar en el programa. Sin embargo, las comparaciones antes-después consideran el contrafactual estimado ($Y \mid P = 0$) como el resultado para el grupo de tratamiento antes de que comience la intervención. Básicamente, esta comparación supone que si el programa no hubiera existido, el resultado (Y) para los participantes del programa habría sido igual a su situación antes del programa. Lo cierto es que en la mayoría de los programas implementados durante meses o años este supuesto no puede sostenerse.

A modo de ejemplo se aborda la evaluación de un programa de microfinanzas para agricultores pobres. El programa ofrece micropréstamos a los agricultores, lo que les permite comprar fertilizantes y aumentar su

producción de arroz. El año previo al inicio del programa, los agricultores cosecharon un promedio de 1.000 kilos de arroz por hectárea (el punto B en el gráfico 3.3). Un año después del lanzamiento del sistema de microfinanzas, la producción de arroz subió a 1.100 kilos por hectárea (el punto A en el gráfico 3.3). Si se tratara de evaluar el impacto sobre la base de una comparación antes-después, se utilizaría el resultado de antes del programa como una estimación del contrafactual. Al aplicar la misma fórmula básica de la evaluación de impacto, se llegaría a la conclusión de que el programa ha incrementado la producción de arroz en 100 kilos por hectárea (A-B).

Ahora bien, si las lluvias fueron normales en el año previo al lanzamiento del programa y luego, durante el año en que se implementó el programa, se produjo una sequía, la producción promedio de los agricultores sin el sistema de micropréstamos será probablemente menor que B, y se situará, por ejemplo, en el nivel D debido a la sequía. En ese caso, el verdadero impacto del programa sería A-D, que es mayor que los 100 kilos estimados a partir de la comparación antes-después. En cambio, si las lluvias hubieran mejorado entre los dos años, el contrafactual de la producción de arroz podría haberse situado en el nivel C. Entonces el verdadero impacto

Gráfico 3.3 Estimaciones antes-después de un programa de microfinanzas

Nota: Δ = Cambio en la cosecha de arroz (kg); ha = hectáreas; kg = kilogramos.

del programa habría sido inferior a 100 kilos. En otras palabras, si los análisis de impacto no pueden dar cuenta de las lluvias y de *todos los demás factores* capaces de influir en la producción de arroz a lo largo del tiempo, es imposible calcular el verdadero impacto del programa mediante una comparación antes-después.

En el ejemplo anterior de las microfinanzas, las lluvias constituyen uno de varios factores externos que pueden influir en el resultado de interés del programa (la producción de arroz) durante su vigencia. De la misma manera, muchos resultados que los programas de desarrollo pretenden mejorar, como los ingresos, la productividad, la salud o la educación, están afectados por un conjunto de factores en el transcurso del tiempo. Por ese motivo, el resultado de referencia casi nunca es una buena estimación del contrafactual y se lo considera un contrafactual "falso".

Evaluación de impacto del HISP: una comparación de resultados antes-después

Como se recordará, el HISP es un programa nuevo en el país, que subsidia los seguros de salud en los hogares rurales pobres, y este seguro cubre los gastos relacionados con la atención de salud y los medicamentos para quienes se inscriben en él. El objetivo del HISP es reducir lo que los hogares pobres gastan en atención primaria de salud y medicamentos y, eventualmente, mejorar los resultados de salud. Aunque se podrían contemplar numerosos indicadores de resultados para la evaluación del programa, al gobierno le interesa, en particular, el análisis de los efectos del HISP en los gastos directos en salud per cápita anuales (de ahora en adelante denominados "gasto en salud").

Si el HISP se ampliara a todo el país representaría un alto porcentaje del presupuesto nacional, hasta el 1,5% del producto interno bruto (PIB), según algunas estimaciones. Además, hay otras complejidades administrativas y logísticas importantes que intervienen en la gestión de un programa de estas características. Por eso, en los niveles más altos del gobierno se ha tomado la decisión de introducir el HISP primero como programa piloto y, más tarde, según los resultados de la fase inicial, aumentar su escala de manera progresiva a lo largo del tiempo. Sobre la base de los resultados de los análisis financieros y de costo-beneficio, la presidenta y su gabinete han anunciado que para que el HISP sea viable y se pueda extender a todo el país es preciso que los gastos en salud per cápita anuales de los hogares rurales pobres disminuyan en al menos US$9 en promedio, en comparación con lo que habrían gastado en ausencia del programa, y se debe lograr esta meta en un plazo de dos años.

Finalmente, durante la fase piloto inicial el HISP se implementa en 100 pueblos rurales. Justo antes del comienzo del programa, el gobierno contrata una empresa de encuestas para que realice un sondeo de línea de base en los 4.959 hogares de estos pueblos. La encuesta recopila información detallada sobre cada hogar, que incluye su composición demográfica, sus activos, su acceso a los servicios sanitarios y su gasto en salud durante el año anterior. Poco después de la encuesta de línea de base, el HISP llega a los 100 pueblos piloto con un gran despliegue de publicidad, que abarca actividades comunitarias y otras campañas promocionales para animar a los pobladores a inscribirse.

De los 4.959 hogares que contiene la muestra de línea de base, un total de 2.907 se inscriben en el HISP y el programa funciona con éxito durante los dos años siguientes. Todas las clínicas y farmacias que sirven los 100 pueblos aceptan a los pacientes con el sistema de seguro y las encuestas muestran que la mayoría de los hogares inscritos expresa satisfacción con el programa. Al final del período piloto de dos años se recopila una segunda ronda de datos de evaluación en la misma muestra de 4.959 hogares.[3]

La presidenta y el ministro de Salud le han encargado a un equipo que supervise la evaluación de impacto del HISP y este debe recomendarle al gobierno si es conveniente ampliar el programa al nivel nacional o no. La pregunta de evaluación de impacto de interés es: ¿cuál es el efecto del Programa de Subsidios de Seguros de Salud en los gastos directos en salud de los hogares pobres? Es preciso recordar que hay mucho en juego. Si se llega a la conclusión de que el HISP reduce los gastos en salud en al menos US$10, se extenderá a todo el país. Si el programa no alcanza ese objetivo, la decisión será que no se amplíe.

El primer consultor "experto" señala que para estimar el impacto del HISP hay que calcular el cambio en los gastos en salud a lo largo del tiempo en los hogares que se inscribieron. El consultor sostiene que debido a que el HISP cubre todos los gastos de salud, cualquier reducción de los gastos durante la vigencia del programa debe ser atribuida al efecto del HISP. A partir del subconjunto de los hogares inscritos, el experto estima los gastos promedio en salud antes de la implementación del programa y luego de la ejecución del mismo, dos años después. En otras palabras, lleva a cabo una comparación antes-después (los resultados se recogen en el cuadro 3.1) Se observa que el grupo de tratamiento redujo sus gastos directos en salud en US$6,65, al pasar de US$14,49 antes de la introducción del HISP a US$7,84 dos años más tarde. Como indica el valor *t*-estadístico, la diferencia entre gastos en salud antes y después del programa es *estadísticamente significativa*.[4] Esto significa que se ha

Cuadro 3.1 Evaluación del HISP según comparación antes-después

	Después	Antes	Diferencia	*t*-estadístico
Gasto en salud de los hogares (en US$)	7,84	14,49	–6,65**	–39,76

** Significativo al nivel del 1%.

Cuadro 3.2 Impacto del HISP según comparación antes-después (análisis de regresión)

	Regresión lineal	Regresión lineal multivariante
Impacto estimado en el gasto en salud de los hogares (en US$)	–6,65** (0,23)	–6,71** (0,23)

Nota: Los errores estándar están entre paréntesis.

** Significativo al nivel del 1%.

encontrado evidencia sólida contra quienes sostienen que la verdadera diferencia entre los gastos antes y después de la intervención es cero.

Aunque la comparación antes-después es para el mismo grupo de hogares, es preciso establecer si otras circunstancias cambiaron en estos hogares a lo largo de los últimos dos años, influyendo en sus gastos en salud. Por ejemplo, hay nuevos medicamentos que se encuentran disponibles desde hace poco tiempo. Además, la reducción de los gastos en salud podría haber sido provocada por la crisis financiera que el país ha vivido recientemente. Para abordar algunas de estas cuestiones, el consultor lleva a cabo un análisis de regresión más sofisticado que intentará controlar por estos otros factores.

El análisis de regresión emplea las estadísticas para analizar las relaciones entre una variable dependiente (la variable que será explicada) y variables explicativas (los resultados se recogen en el cuadro 3.2). Una regresión lineal es la forma más sencilla de hacerlo: los gastos en salud son la variable dependiente y solo hay una variable explicativa, un indicador binario (0-1) que asume el valor 0 si la observación es de línea de base y 1 si la observación corresponde al seguimiento.

En tanto, una regresión lineal multivariante añade variables explicativas para *controlar por*, o *mantener constantes*, otras características que se observan para los hogares en la muestra, que incluyen indicadores de riqueza (activos), composición del hogar, etcétera.[5]

Se observa que el resultado de la regresión lineal es equivalente a la simple diferencia antes y después en los gastos promedio en salud que muestra el cuadro 3.1 (una reducción de US$6,65 en los gastos en salud). Una vez que se utiliza una regresión lineal multivariante para controlar por otros factores disponibles en los datos, se vuelve a obtener un resultado similar: una disminución de US$6,71 en los gastos en salud.

Pregunta HISP 1

A. ¿La comparación antes-después controla por todos los factores que influyen en los gastos en salud a lo largo del tiempo?

B. Sobre la base de los resultados producidos por el análisis antes-después, ¿debería ampliarse el HISP a nivel nacional?

Contrafactual falso 2: comparación entre los grupos de inscritos y no inscritos (autoseleccionados)

La comparación entre un grupo de individuos que se inscriben voluntariamente en un programa y un grupo de individuos que *elige* no participar es otro enfoque riesgoso de la evaluación de impacto. Un grupo de comparación que se autoselecciona para no participar en un programa será otro contrafactual falso. La selección se produce cuando la participación en el programa se basa en las preferencias, decisiones o características no observables de los participantes potenciales.

A modo de ejemplo se puede pensar en un programa de orientación profesional para los jóvenes desempleados. Dos años después de su lanzamiento, una evaluación intenta estimar su impacto en los ingresos a partir de la comparación de los ingresos promedio de un grupo de jóvenes que decidieron inscribirse en el programa con los de un grupo de jóvenes que, a pesar de ser elegibles, decidieron no inscribirse. Si los resultados demostraran que los jóvenes que eligieron inscribirse en el programa ganan el doble de los que decidieron no hacerlo, ¿cómo debería interpretarse este hallazgo? En este caso, el contrafactual se estima sobre la base de los ingresos de quienes eligieron no inscribirse en el programa. Sin embargo, es probable que los dos grupos sean, en esencia, diferentes. Aquellos individuos que decidieron participar pueden estar muy motivados para mejorar sus vidas y quizá esperen un retorno alto de la formación. Mientras que los que decidieron no inscribirse tal vez son jóvenes desanimados que no esperan beneficiarse de este tipo de programas. Es factible que estos dos grupos

obtengan resultados bastante distintos en el mercado laboral y consigan ingresos diferentes incluso sin el programa de formación profesional.

Los mismos problemas surgen cuando la admisión en un programa se basa en preferencias no observadas de los administradores del programa. Por ejemplo, si los administradores del programa basan la admisión y la inscripción en una entrevista. Puede que los admitidos en el programa sean aquellos en quienes los administradores ven una buena probabilidad de beneficiarse del programa. Tal vez los no admitidos pueden mostrar menos motivación en la entrevista, tener calificaciones más bajas o sencillamente carecer de destrezas en una entrevista. Como se señaló en el caso anterior, es probable que estos dos grupos de jóvenes obtengan ingresos diferentes en el mercado laboral incluso sin un programa de formación profesional.

Por lo tanto, el grupo que no se inscribió no proporciona una buena estimación del contrafactual ya que la observación de una diferencia en los ingresos entre los dos grupos no permite determinar si se debe al programa de formación o a los contrastes subyacentes entre los dos grupos en motivación, destrezas y otros factores. Así, el hecho de que individuos menos motivados o menos cualificados no se hayan inscrito en el programa de formación genera un sesgo en la evaluación de impacto del programa.[6] Este sesgo se llama *sesgo de selección*. En términos más generales, el sesgo de selección se produce cuando los motivos por los que un individuo participa en un programa están correlacionados con los resultados, incluso en ausencia del programa. Asegurarse de que el impacto estimado esté libre de sesgos de selección es uno de los principales objetivos de cualquier evaluación de impacto, y plantea importantes dificultades. En este ejemplo, si los jóvenes que se inscribieron en la formación profesional hubiesen tenido ingresos más altos incluso en ausencia del programa, el sesgo de selección sería positivo; en otras palabras, se sobreestimaría el impacto del programa de formación profesional al atribuirle los ingresos más altos que los participantes habrían tenido de todas maneras.

Concepto clave

El sesgo de selección se produce cuando los motivos por los que un individuo participa en un programa están correlacionados con los resultados. Asegurarse de que el impacto estimado esté libre de sesgos de selección es uno de los principales objetivos en cualquier evaluación de impacto y plantea importantes dificultades.

Evaluación del impacto del HISP: comparación entre hogares inscritos y no inscritos

Después de haber reflexionado de forma más detenida sobre la comparación antes-después el equipo de evaluación llega a la conclusión de que todavía hay numerosos factores que pueden explicar parte del cambio en los gastos en salud a lo largo del tiempo (concretamente, al ministerio de Finanzas le preocupa que una reciente crisis financiera haya afectado los ingresos de los hogares, y puede que explique el cambio observado en los gastos en salud).

Otro consultor sugiere que sería más adecuado estimar el contrafactual en el período posterior a la intervención, es decir, dos años después del comienzo del programa. El consultor señala que de los 4.959 hogares contenidos en la muestra, solo 2.907 se inscribieron en el programa, de modo que alrededor del 41% de los hogares sigue sin cobertura del HISP. El consultor sostiene que todos los hogares de los 100 pueblos piloto cumplían las condiciones para inscribirse. Estos hogares comparten las mismas clínicas de salud y están sujetos a los mismos precios locales de los productos farmacéuticos. Además, la mayoría de los integrantes de esos hogares trabaja en actividades económicas similares. El consultor opina que, en estas circunstancias, los resultados del grupo no inscrito después de la intervención podrían servir para estimar el resultado contrafactual del grupo inscrito en el HISP. Por lo tanto, decide calcular los gastos promedio en salud en el período posterior a la intervención, tanto para los hogares que se inscribieron en el programa como para los que no lo hicieron (los resultados se recogen en el cuadro 3.3). Utilizando los gastos promedio de salud de los hogares no inscritos como la estimación del contrafactual, el consultor llega a la conclusión de que el programa ha reducido los gastos promedio de salud en casi US$14,46.

Ahora bien, los hogares que decidieron no inscribirse en el programa ¿pueden ser sistemáticamente diferentes de los que sí lo hicieron? Quizá los hogares que se inscribieron en el HISP tenían mayores gastos en salud o eran personas con más información acerca del programa o más atentas a la salud de su familia. Otra posibilidad es que tal vez los hogares que se inscribieron eran más pobres, en promedio, que los que no se inscribieron, ya que el HISP tenía como objetivo los hogares pobres. El consultor asegura que el análisis de regresión puede controlar por estas diferencias potenciales entre los dos grupos. Por lo tanto, realiza otra regresión multivariante que controla por todas las características del hogar que puede encontrar en la base de datos, y estima el impacto del programa como se muestra en el cuadro 3.4.

Cuadro 3.3 Evaluación del HISP según comparación inscritos-no inscritos (comparación de medias)

	Inscritos	No inscritos	Diferencia	*t*-estadístico
Gasto en salud de los hogares (en US$)	7,84	22,30	–14,46**	–49,08

** Significativo al nivel del 1%.

Cuadro 3.4 Evaluación del HISP según comparación inscritos-no inscritos (análisis de regresión)

	Regresión lineal	Regresión lineal multivariante
Impacto estimado sobre el gasto en salud de los hogares (en US$)	–14,46** (0,33)	–9,98** (0,29)

Nota: Los errores estándar están entre paréntesis.

** Significativo al nivel del 1%.

Con una simple regresión lineal de los gastos en salud en una variable indicativa de si un hogar se inscribió o no en el programa, es posible encontrar un impacto estimado de US$ -14,46, es decir, que el programa ha disminuido el promedio de gastos de salud en US$14,46. Sin embargo, cuando se controla por todas las demás características de los datos, se estima que el programa ha reducido los gastos en salud en US$ 9,98 al año.

Pregunta HISP 2

A. ¿Este análisis controla por todos los factores que determinan las diferencias en gastos en salud entre los dos grupos?

B. Sobre la base de los resultados producidos por el método de inscripción-no inscripción, ¿debería ampliarse el HISP al nivel nacional?

Recursos adicionales

- Para material complementario del libro y para otros hipervínculos de recursos, se recomienda ver el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).

Notas

1. Usamos el Modelo Causal de Rubin como marco para la inferencia causal.
2. Esta condición se relajará en algunos métodos de evaluación de impacto que, en cambio, requerirán que el cambio promedio en los resultados (tendencias) sea el mismo en ausencia del programa.
3. Se supone que ningún hogar ha dejado la muestra en esos dos años (hay un desgaste cero de la muestra). Este no es un supuesto realista en la mayoría de las encuestas de hogares. En la práctica, a veces no se puede hacer un seguimiento de

las familias que se mudan en su nueva localidad, y algunos hogares se rompen o dejan de existir del todo.

4. Cabe destacar que un *t*-estadístico *(t-stat)* de 1,96 o más (en valor absoluto) es "estadísticamente significativo" en un nivel del 5%.
5. Para más información sobre el análisis multivariante, véase el manual técnico en línea del sitio web de la Evaluación de Impacto en la Práctica (www.worldbank .org/ieinpractice).
6. Otro ejemplo: si los jóvenes que esperan beneficiarse de un modo considerable del sistema de formación también tienen más probabilidades de inscribirse (tal vez porque esperan tener salarios más altos con la formación) compararlos con un grupo de jóvenes que espera menores retornos y que no se inscribe arrojará una estimación de impacto sesgada.

Referencias bibliográficas

Imbens, G. y D. Rubin. 2008. "Rubin Causal Model." En: S. N. Durlauf y L. E. Blume (eds.), *The New Palgrave Dictionary of Economics* (2da. edición). Nueva York: Palgrave.

Rubin, D. 1974. "Estimating Causal Effects of Treatments in Randomized and Non- Randomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

CAPÍTULO 4

La asignación aleatoria

La evaluación de programas basados en reglas de asignación

Después de analizar dos estimaciones "falsificadas" del contrafactual que se utilizan habitualmente, aunque con un alto riesgo de sesgo –las comparaciones antes-después y las comparaciones inscritos-no inscritos–, a continuación se presentará un conjunto de métodos que se pueden aplicar para estimar con mayor precisión los impactos de un programa. Sin embargo, esas estimaciones no siempre son tan fáciles como puede parecer a primera vista. La mayoría de los programas se diseña y luego se implementa en un entorno complejo y cambiante, donde diversos factores pueden influir en los resultados tanto de los participantes del programa como de aquellos que no participan. Las sequías, los terremotos, las recesiones, los cambios de gobierno y los vaivenes de las políticas nacional e internacional forman parte del mundo real. En una evaluación, se debe asegurar que la estimación del impacto del programa siga siendo válida a pesar de esta diversidad de factores.

Las reglas de un programa para seleccionar a los participantes constituirán el parámetro clave para determinar el método de la evaluación de impacto. Creemos que, en la mayoría de los casos, los métodos de evaluación deben intentar adaptarse al contexto de las reglas operativas de un programa (con unos pocos ajustes por aquí y por allá) y no al contrario. Sin embargo, también se parte de la premisa de que *todos los programas*

sociales deben tener reglas de asignación justas y transparentes. Una de las reglas más justas y transparentes para asignar recursos escasos entre poblaciones que los merecen de igual manera consiste en que todos aquellos que sean elegibles tengan la misma oportunidad de participar en el programa. Una manera sencilla de conseguirlo es mediante un sorteo.

En este capítulo, se analizará un método que se asemeja a un sorteo y que decide quién participa en un programa en un determinado momento y quién no: el *método de asignación aleatoria*, también conocido como ensayo aleatorio controlado (RCT, por sus siglas en inglés, *randomized control trial*). Este método no solo proporciona a los administradores del programa una regla imparcial y transparente para asignar recursos escasos *entre poblaciones igualmente merecedoras de ellos*, sino que también representa el método más sólido para evaluar el impacto de un programa. Por eso, la aplicación del mismo para evaluar los impactos de los programas ha aumentado de manera considerable en los últimos años.

La asignación aleatoria del tratamiento

Cuando se asigna de forma aleatoria a los beneficiarios de un programa –es decir, mediante sorteo– entre una población elegible numerosa, se puede generar una estimación robusta del contrafactual. La *asignación aleatoria* del tratamiento se considera la regla de oro de la evaluación de impacto. Utiliza un proceso aleatorio, o el azar, para decidir a quién se le concederá acceso al programa y a quién no.[1] En la asignación aleatoria, todas las unidades elegibles (por ejemplo, una persona, un hogar, una empresa, un hospital, una escuela o una comunidad) tienen la misma probabilidad de ser seleccionadas para un programa.[2]

Antes de ver cómo se implementa la asignación aleatoria en la práctica y por qué genera una estimación sólida del contrafactual, conviene dedicar un momento a pensar por qué la asignación aleatoria también es una manera justa y transparente de asignar los escasos recursos de un programa. Una vez que se ha definido una población objetivo (por ejemplo, hogares por debajo de la línea de la pobreza, niños menores de 5 años o caminos en zonas rurales en el norte del país), la asignación aleatoria es una regla de asignación justa porque permite que los administradores del programa se aseguren de que todas las unidades elegibles tengan la misma probabilidad de participar en el programa y de que el programa no sea asignado con criterios arbitrarios o subjetivos, ni por cuestiones de favoritismo u otras prácticas injustas. Cuando se produce un exceso de demanda de un programa, la asignación aleatoria es una regla que los administradores del mismo pueden explicar fácilmente, que todas las partes interesadas pueden entender y que

se considera justa en numerosas circunstancias. Además, cuando el proceso de asignación se lleva a cabo de modo abierto y transparente, no es fácil manipularlo y, por lo tanto, protege a los administradores del programa de posibles acusaciones de favoritismo o corrupción. Por lo tanto, como mecanismo de asignación, la asignación aleatoria tiene sus propios méritos, que van mucho más allá de su utilidad como instrumento de evaluación de impacto. De hecho, diversos programas utilizan de manera rutinaria los sorteos como una forma de seleccionar a los participantes del conjunto de individuos elegibles, sobre todo debido a sus ventajas administrativas y de gobernanza.[3] El recuadro 4.1 presenta dos casos de este tipo en África.

Recuadro 4.1: La asignación aleatoria como un valioso instrumento operativo

La asignación aleatoria puede ser una regla útil para asignar los beneficios de un programa, incluso fuera del contexto de una evaluación de impacto. Los siguientes dos casos de África ilustran por qué.

En Costa de Marfil, después de un período de crisis, el gobierno introdujo un programa de empleo temporal inicialmente dirigido a los ex combatientes, que luego se amplió a la juventud en términos más generales. El programa ofrecía a los jóvenes oportunidades de empleo de corto plazo, sobre todo limpiando o rehabilitando caminos, a través de la agencia nacional de vialidad. Se invitó a los jóvenes en las municipalidades que participaban a que se inscribieran. Dado el atractivo de los beneficios, fueron muchos más los jóvenes que postularon que las plazas disponibles. Para encontrar una manera transparente y justa de asignar los beneficios entre los postulantes, los administradores del programa recurrieron a un proceso de sorteo público. Una vez que la inscripción se cerró y se conocía el número de postulantes (por ejemplo, *N*) en una localidad, se organizaba un sorteo público. Se convocaba a todos los postulantes a un lugar público, y se introducían en una caja pequeños trozos de papel con números que iban de 1 a *N*. Después, se llamaba a los postulantes uno por uno para que sacaran un número de la caja delante de todos los demás participantes. Una vez que se sacaba el número, se leía en voz alta. Después de que se había llamado a todos los postulantes, se verificaban uno por uno los números que quedaban en la caja para cerciorarse de que correspondieran a participantes que no habían asistido al sorteo. Si había *N* plazas disponibles en el programa, se seleccionaba a aquellos postulantes que habían sacado los números más bajos. El proceso de sorteo se organizó por separado para hombres y mujeres. Fue bien aceptado por los participantes, y contribuyó a dar una imagen de imparcialidad y transparencia al programa en un contexto posterior al conflicto, marcado por tensiones sociales. Después de varios años en marcha, los investigadores utilizaron esta regla de selección, ya integrada en el funcionamiento del programa, para emprender su evaluación de impacto.

Continúa en la página siguiente.

Recuadro 4.1: La asignación aleatoria como un valioso instrumento operativo *(continúa)*

En Níger, el gobierno comenzó a implementar en 2011 un proyecto de red de protección nacional, con el apoyo del Banco Mundial. Níger es uno de los países más pobres del mundo y la población de hogares pobres que merecían el programa superó por mucho los beneficios disponibles durante los primeros años de funcionamiento. Los administradores del programa contaban con un sistema de selección geográfica para definir en qué departamentos y comunas se implementaría primero el programa de transferencias de efectivo. Esto se podía hacer porque había datos para determinar la pobreza relativa o el estado de vulnerabilidad de los diversos departamentos o comunas. Sin embargo, dentro de las comunas, había muy pocos datos disponibles para evaluar qué pueblos eran más merecedores que otros sobre la base de criterios objetivos. Así, para la primera fase del proyecto, los administradores del programa decidieron utilizar sorteos públicos de modo de seleccionar a los pueblos beneficiarios dentro de las comunas definidas como objetivo. Esta decisión se llevó a cabo en parte porque los datos disponibles para priorizar de manera objetiva a los pueblos eran limitados, y en parte porque en el proyecto se estaba incorporando una evaluación de impacto. En los sorteos públicos se invitaba a todos los responsables de los pueblos al centro municipal, se escribían los nombres de sus pueblos en un trozo de papel, y se introducían en una caja. Después, un niño sacaba al azar los papeles que correspondían a los pueblos beneficiarios hasta que se llenaban las cuotas. El procedimiento se utilizó por separado para pueblos sedentarios y pueblos nómadas a fin de asegurar la representación de cada grupo. (Después de seleccionar a los pueblos, se implementó un mecanismo de definición de objetivos a nivel de los hogares para identificar a los hogares más pobres, que posteriormente fueron inscritos como beneficiarios.) La transparencia y la imparcialidad del sorteo público fueron muy apreciadas por las autoridades locales y de los pueblos, y por los administradores del programa, de tal manera que el proceso de sorteo público siguió siendo utilizado en el segundo y tercer ciclo del proyecto para seleccionar más de 1.000 pueblos en todo el país. Aunque el sorteo público no fue necesario para realizar una evaluación de impacto en ese momento, su valor como instrumento operativo transparente, justo y ampliamente aceptado para asignar beneficios entre poblaciones que los merecían por igual, justificó la continuidad de su uso ante los administradores del programa y las autoridades locales.

Fuentes: Bertrand et al. (2016); Premand, Barry y Smitz (2016).

La asignación aleatoria a menudo puede derivarse de las reglas operativas de un programa. En numerosos programas, la población de posibles participantes –es decir, el conjunto de individuos que el programa quisiera servir– es mayor que el número de participantes a los que el programa se puede realmente dirigir en un determinado momento. Por ejemplo, en un solo año un programa educativo puede proporcionar materiales escolares a 500 escuelas de un total de 1.000 escuelas elegibles en el país.

O un programa de mejora de caminos rurales puede tener el objetivo de pavimentar 250 caminos rurales, aunque haya cientos de caminos más que el programa desearía mejorar. O un programa de empleo para jóvenes puede tener la meta de llegar a 2.000 jóvenes desempleados en su primer año de funcionamiento, aunque haya decenas de miles de jóvenes en dichas circunstancias que el programa quisiera servir. Por diversos motivos, puede que los programas no logren alcanzar al conjunto de la población de interés. Las limitaciones presupuestarias pueden impedir que los administradores ofrezcan el programa a todas las unidades elegibles desde el comienzo. Aunque haya presupuestos disponibles para cubrir un gran número de participantes, las limitaciones de capacidad a veces impedirán que un programa pueda ser implementado para todos al mismo tiempo. Por ejemplo, en el caso del programa de formación profesional para jóvenes, la cantidad de jóvenes desempleados que desean obtener una formación profesional puede ser superior al número de plazas disponibles en las escuelas técnicas durante el primer año del programa, y eso puede restringir el número de alumnos que pueden matricularse.

Cuando la población de participantes elegibles es superior al número de plazas disponibles en el programa, alguien debe tomar la decisión de quién participará y quién no. En otras palabras, los administradores del programa deben definir un mecanismo de selección para asignar los servicios del mismo. El programa podría funcionar según un principio de orden de llegada, o basarse en características observables (por ejemplo, atendiendo primero las zonas más pobres); la selección también podría basarse en características no observables (por ejemplo, dejar que los individuos se inscriban a partir de sus propias motivaciones y conocimientos), o se podría recurrir a un sorteo. Incluso en contextos donde es posible clasificar a los participantes potenciales en función de la necesidad, puede que convenga asignar parte de los beneficios mediante un sorteo. Por ejemplo, piénsese en un programa que tiene como objetivo al 20% más pobre de los hogares sobre la base de una medida del ingreso. Si el ingreso solo se puede medir de forma imperfecta, el programa podría usar esta medida para incluir a todos los participantes potenciales que se identifican como "de extrema pobreza" (por ejemplo, el 15% inferior). Sin embargo, dado que el ingreso se mide de manera imperfecta, los hogares que se encuentren justo por debajo del umbral de elegibilidad en el percentil 20º, en la realidad pueden ser elegibles o no serlo (si se pudiera medir el verdadero ingreso), mientras que los hogares que se hallen justo por encima del percentil 20º también pueden ser elegibles o no. En este contexto, utilizar un sorteo para determinar qué hogares serían beneficiarios en torno al percentil 20º (por ejemplo, entre los percentiles 15º y 25º de la distribución del ingreso) podría ser una manera justa de asignar los beneficios en este grupo de hogares.

¿Por qué la asignación aleatoria produce una excelente estimación del contrafactual?

Como ya se ha visto, el grupo de comparación ideal sería lo más similar posible al grupo de tratamiento en todos los sentidos, excepto con respecto a su participación en el programa que se evalúa. Cuando se asignan unidades de manera aleatoria a los grupos de tratamiento y de comparación, ese proceso de asignación aleatoria producirá dos grupos que tienen una alta probabilidad de ser estadísticamente idénticos, siempre que el número de unidades potenciales a las que se aplica el proceso de asignación aleatoria sea suficientemente grande. Concretamente, con un gran número de unidades el proceso de asignación aleatoria producirá grupos que tienen *promedios estadísticamente equivalentes en todas sus características*.[4]

Concepto clave

En la asignación aleatoria, cada unidad elegible tiene la misma probabilidad de ser seleccionada para el tratamiento, de modo que se asegura la equivalencia entre los grupos de tratamiento y comparación tanto en las características observables como en las no observables.

El gráfico 4.1 ilustra por qué la asignación aleatoria produce un grupo de comparación estadísticamente equivalente al grupo de tratamiento. Supóngase que la población de unidades elegibles (el conjunto de participantes potenciales, o la población de interés para la evaluación) consiste en más de 1.000 personas. Entonces, se asigna aleatoriamente la mitad al grupo de tratamiento y la otra mitad al grupo de comparación. Por ejemplo, se escriben los nombres de las 1.000 personas en trozos de papel individuales, se mezclan todos los trozos en una caja, y luego se le pide a alguien que extraiga a ciegas 500 nombres. Si los primeros 500 nombres constituyen el grupo de tratamiento, entonces tendríamos un grupo de tratamiento asignado de forma aleatoria (los primeros 500 números extraídos) y un grupo de comparación asignado también de manera aleatoria (los 500 nombres que quedaron en la caja).

Ahora supóngase que el 40% de las 1.000 personas originales eran mujeres. Dado que los nombres se han asignado al azar, de los 500 nombres que

Gráfico 4.1 Características de los grupos bajo tratamiento con asignación aleatoria

se sacaron de la caja, alrededor del 40% serán también mujeres. Si entre las 1.000 personas, el 20% tenía los ojos azules, entonces casi el 20% de ellas en los grupos de tratamiento y de comparación también deberían tener los ojos azules. En general, si la población de unidades elegibles es lo suficientemente grande, el mecanismo de asignación aleatoria asegura que cualquier característica de la población se transfiera tanto al grupo de tratamiento como al grupo de comparación. Del mismo modo que las características observables, como el sexo o el color de los ojos de un individuo, se transfieren tanto al grupo de tratamiento como al de comparación, es lógico pensar que las características que son más difíciles de observar (*variables no observables*), como la motivación, las preferencias u otros rasgos de la personalidad que son complejos de medir, también se aplicarán por igual al grupo de tratamiento y al de comparación. Así, los grupos de tratamiento y comparación generados a través de la asignación aleatoria serán similares no solo en cuanto a sus características observables sino también en relación con las no observables. Tener dos grupos similares en todos los aspectos asegura que la estimación del contrafactual se aproxime al valor verdadero del resultado en ausencia de tratamiento, y que una vez que el programa se haya implementado, las estimaciones de impacto no sufrirán un sesgo de selección.

Cuando una evaluación utiliza la asignación aleatoria para generar los grupos de tratamiento y de comparación, en teoría, el proceso debería producir dos grupos equivalentes, siempre que se cuente con un número de unidades lo suficientemente grande. Con los datos de línea de base de la muestra de evaluación con la que se cuente, se podrá comprobar empíricamente este supuesto y verificar que, de hecho, no hay diferencias sistemáticas en las características observables entre los grupos de tratamiento y de comparación antes del inicio del programa. Luego, si después de lanzar el programa se observan diferencias en los resultados entre los grupos de tratamiento y comparación, sabremos que esas diferencias se deben únicamente a la incidencia del programa, dado que los dos grupos eran idénticos en la línea de base, antes del inicio del programa, y que están expuestos a los mismos factores externos a lo largo del tiempo. En este sentido, el grupo de comparación *contiene* todos los factores que también pueden explicar el resultado de interés.

Para estimar el impacto de un programa bajo la asignación aleatoria, se debe observar la diferencia entre el resultado bajo tratamiento (el resultado medio del grupo de tratamiento asignado de forma aleatoria) y nuestra estimación del contrafactual (el resultado medio del grupo de comparación asignado de manera aleatoria). Así, podemos confiar en que nuestro impacto estimado constituye el verdadero impacto del programa, puesto que se han eliminado todos los factores observados y no observados que, de otra manera, podrían explicar la diferencia en los resultados.

En los recuadros 4.2 a 4.6, se analizan las aplicaciones de la asignación aleatoria en el mundo real para evaluar el impacto de diversas intervenciones en todo el mundo.

En el gráfico 4.1 se presupone que todas las unidades de la población elegible serán asignadas ya sea al grupo de tratamiento o al grupo de comparación. Sin embargo, en algunos casos no es necesario incluir todas las unidades en la evaluación. Por ejemplo, si la población de unidades elegibles es de 1 millón de madres y se quiere evaluar la efectividad de los bonos en efectivo en la probabilidad de que estas madres vacunen a sus hijos, bastará con seleccionar una muestra aleatoria representativa de, por ejemplo, 1.000 madres y asignar a esas 1.000 madres ya sea al grupo de tratamiento o de comparación. El gráfico 4.2 ilustra este proceso. Según la misma lógica que

Recuadro 4.2: La asignación aleatoria como regla de selección de un programa: las transferencias condicionadas y la educación en México

El programa Progresa, actualmente denominado "Prospera", proporciona transferencias en efectivo a las madres pobres de zonas rurales de México con la condición de que matriculen a sus hijos en la escuela y acudan regularmente a chequeos de salud (véase el recuadro 1.1 en el capítulo 1). Las transferencias en efectivo para niños de entre tercero y noveno grado equivalen a alrededor del 50% al 75% del costo privado de la escolarización y están garantizadas durante tres años. Las comunidades y los hogares elegibles para el programa se determinaron sobre la base del índice de pobreza creado a partir de los datos del censo y de la recopilación de datos básicos. Debido a la necesidad de desplegar en fases el programa social de gran escala, cerca de las dos terceras partes de las localidades (314 de 495) fueron asignadas aleatoriamente para ser beneficiarias del programa en los dos primeros años, y las otras 181 sirvieron como grupo de comparación antes de incorporarse al programa en el tercer año.

Sobre la base de la asignación aleatoria, Schultz (2004) encontró un aumento promedio de un 3,4% en la inscripción de todos los alumnos de primero a octavo grados, y el mayor incremento, de un 14,8%,[a] se observó entre las niñas que habían terminado sexto grado. El probable motivo de este último resultado es que la tasa de niñas que tienden a abandonar la escuela crece a medida que estas son mayores; por ello, con el fin de que permanecieran en la escuela después de los cursos de primaria, las niñas recibieron una transferencia ligeramente superior. Estos impactos de corto plazo luego fueron extrapolados para predecir el impacto a más largo plazo del programa Progresa en la escolarización y en los ingresos a lo largo de la vida.

Fuente: Schultz (2004).
a. Para ser precisos, Schultz combinó la asignación aleatoria con el método de diferencias en diferencias que se analiza en el capítulo 7.

Recuadro 4.3: Asignación aleatoria de donaciones para mejorar las perspectivas de empleo juvenil en el norte de Uganda

En 2005 el gobierno de Uganda comenzó un programa destinado a disminuir el desempleo juvenil y a promover la estabilidad social en el norte del país, sacudido por conflictos. El Programa de Oportunidades para los Jóvenes invitó a grupos de jóvenes adultos a presentar propuestas de ayuda para actividades empresariales y formación profesional. Se presentaron miles de propuestas, pero el gobierno solo podía financiar unas cuantas centenas.

Aprovechando la alta demanda del programa, los evaluadores trabajaron con el gobierno para asignar de forma aleatoria cuáles serían los grupos que recibirían financiamiento. El gobierno central pidió a los gobiernos de los distritos que presentaran más del doble de propuestas de las que podían financiar. Después de una selección, el gobierno elaboró una lista de 535 propuestas elegibles para el programa. Posteriormente, las propuestas fueron asignadas de manera aleatoria a grupos de tratamiento (265 propuestas) o de comparación (270 propuestas).

El monto de la ayuda en el grupo de tratamiento ascendía a un promedio de US$382 por persona. Cuatro años después del desembolso, los jóvenes del grupo de tratamiento tenían más del doble de probabilidades de trabajar en un oficio calificado que los jóvenes del grupo de comparación. También ganaban un 38% más y ostentaban un 57% más de *stock* de capital. Sin embargo, los investigadores no encontraron impacto alguno en la cohesión social ni en las conductas antisociales.

Fuente: Blattman, Fiala y Martínez (2014).

Recuadro 4.4: Asignación aleatoria de intervenciones en abastecimiento de agua y saneamiento en zonas rurales de Bolivia

A partir de 2012, el gobierno boliviano, con apoyo del Banco Interamericano de Desarrollo (BID), implementó una asignación aleatoria de intervenciones en materia de abastecimiento de agua y saneamiento en pequeñas comunidades rurales. En los 24 municipios del país con las mayores necesidades, el programa identificó más de 369 comunidades elegibles para la intervención. Dado que solo había recursos para cubrir 182 comunidades, el programa utilizó la asignación aleatoria para dar a cada comunidad elegible la misma probabilidad de participar. Junto con los gobiernos municipales, los administradores del programa organizaron una serie de eventos donde celebraron sorteos públicos ante la presencia de dirigentes comunitarios, la prensa y la sociedad civil.

Primero, se dividieron las comunidades según el tamaño de la población. Luego, dentro de cada grupo, se obtuvieron al azar los nombres de las comunidades y se

Continúa en la página siguiente.

Recuadro 4.4: Asignación aleatoria de intervenciones en abastecimiento de agua y saneamiento en zonas rurales de Bolivia *(continúa)*

registraron en una lista. Las comunidades que quedaron al comienzo de la lista se asignaron al grupo de tratamiento. Cada concurso fue monitoreado por un notario público independiente, que posteriormente registró y certificó los resultados, lo que concedió un nivel adicional de legitimidad al proceso. En el caso de las comunidades que quedaron fuera del programa, los gobiernos municipales se comprometieron a utilizar la misma lista ordenada de forma aleatoria para asignar un futuro financiamiento después de completar la evaluación. De esta manera, ninguna comunidad quedaría marginada de la intervención debido únicamente a los objetivos de la evaluación, pero existiría un grupo de comparación mientras las limitaciones presupuestarias restringieran el número de proyectos en cada municipalidad.

Fuente: Proyecto Banco Interamericano de Desarrollo Nº BO-L1065, véase http://www.iadb.org/en/projects/project-description-title,1303.html?id=BO-L1065.
Nota: Véase el sorteo público para asignaciones aleatorias en https://vimeo.com/86744573.

Recuadro 4.5: Asignación aleatoria de protección del agua de pozos para mejorar la salud en Kenia

El vínculo entre calidad del agua e impactos en la salud en los países en desarrollo ha sido bien documentado. Sin embargo, el valor sanitario de mejorar la infraestructura cerca de las fuentes de agua es menos evidente. Kremer et al. (2011) midieron los efectos de un programa que proporcionaba tecnología de protección de pozos para mejorar la calidad del agua en Kenia, asignando aleatoriamente los pozos receptores del tratamiento.

Alrededor del 43% de los hogares de las zonas rurales de Kenia occidental obtienen el agua potable de pozos naturales. La tecnología de protección de fuentes de agua aísla la fuente de un pozo para disminuir la contaminación. A partir de 2005, la ONG *International Child Support* (ICS), implementó un programa de protección de pozos en dos distritos de Kenia occidental. Debido a limitaciones financieras y administrativas, ICS decidió ampliar el programa a lo largo de cuatro años. Esto les permitió a los evaluadores utilizar los pozos que todavía no habían recibido tratamiento como grupo de comparación.

De los 200 pozos elegibles, 100 fueron asignados al azar para recibir el tratamiento en los primeros dos años. El estudio observó que la protección de los pozos redujo la contaminación fecal del agua en un 66% y la diarrea infantil entre los usuarios de los pozos en un 25%.

Fuente: Kremer et al. (2011).

Recuadro 4.6: Asignación aleatoria e información a propósito de los riesgos del VIH para reducir el embarazo adolescente en Kenia

En un experimento aleatorio que se realizó en Kenia occidental, Dupas (2011) probó la efectividad de dos diferentes tratamientos de educación sobre el VIH/Sida para reducir conductas sexuales no seguras entre los adolescentes. El primer tratamiento consistió en la formación de profesores en el programa nacional de estudios sobre VIH/Sida, que se centró en la aversión al riesgo y que promovía la abstinencia. El segundo tratamiento, la Campaña de información sobre el riesgo relativo, tenía como objetivo reducir las relaciones sexuales entre hombres mayores y chicas jóvenes proporcionando información sobre las tasas de VIH desagregadas por edad y sexo.

El estudio se llevó a cabo en dos distritos rurales de Kenia, con una muestra de 328 escuelas primarias. Los investigadores asignaron de forma aleatoria 163 escuelas estratificadas por localidad, puntuación de las pruebas y proporción alumnos/sexo, para que recibieran el primer tratamiento. Después, se asignaron de manera aleatoria 71 escuelas al segundo tratamiento estratificando la muestra por su participación en el primer tratamiento. Esto produjo cuatro grupos de escuelas: las escuelas que recibían el primer tratamiento, las que recibían el segundo, las que recibían ambos y las que no recibían ninguno.

La asignación aleatoria de las escuelas garantizaba que no habría diferencias sistemáticas en la información a la que estaban expuestos los alumnos antes de que comenzara el programa. Un año después de la terminación del programa, Dupas observó que la campaña de información sobre el riesgo relativo produjo una disminución del 28% en la probabilidad de que una joven quedara embarazada. En cambio, las escuelas que solo habían sido destinatarias del primer tratamiento no mostraron efecto alguno en el embarazo adolescente.

Fuente: Dupas (2011).

Gráfico 4.2 Muestra aleatoria y asignación aleatoria de tratamiento

se detalló más arriba, la selección de una muestra aleatoria de la población de unidades elegibles para formar la muestra de evaluación conserva las características de la población de las unidades elegibles. Dentro de la muestra, la asignación aleatoria de individuos a los grupos de tratamiento y comparación también conserva dichas características. En el capítulo 15 se abordarán otros aspectos del muestreo.

Validez externa e interna

Los pasos de la selección aleatoria del tratamiento, que ya se han detallado, aseguran tanto la validez interna como externa de las evaluaciones de impacto (gráfico 4.2).

Validez interna significa que el impacto estimado del programa es el impacto libre de todos los demás factores de confusión potenciales (o, en otras palabras, que el grupo de comparación represente una estimación precisa del contrafactual de modo que se estime el verdadero impacto del programa). Hay que recordar que la asignación aleatoria produce un grupo de comparación que es estadísticamente equivalente al grupo de tratamiento en la línea de base, antes de que empiece el programa. Una vez que el programa comienza, el grupo de comparación está expuesto al mismo conjunto de factores externos que el grupo de tratamiento a lo largo del tiempo, con la única excepción del propio programa. Por lo tanto, si aparece cualquier diferencia en los resultados entre los grupos de tratamiento y de comparación, solo puede deberse a la existencia del programa en el grupo de tratamiento. La validez interna de una evaluación de impacto se asegura a través del proceso de *asignación aleatoria del tratamiento*.

Concepto clave
Una evaluación tiene validez interna si proporciona una estimación precisa del contrafactual mediante un grupo de comparación válido.

Validez externa quiere decir que la *muestra* de la evaluación representa con precisión a la población de unidades elegibles. Los resultados de la evaluación se pueden entonces generalizar a la población de unidades elegibles. Se utiliza el *muestreo aleatorio* para asegurar que la muestra de la evaluación refleje adecuadamente la población de unidades elegibles, de modo que los impactos identificados en la muestra de la evaluación pueden extrapolarse a la población.

Concepto clave
Una evaluación tiene validez externa si la muestra de evaluación representa con precisión a la población de unidades elegibles. Los resultados pueden luego generalizarse al conjunto de la población de unidades elegibles.

Nótese que se ha realizado un proceso de selección aleatoria con dos objetivos diferentes: *selección aleatoria* de una muestra (para la validez externa), y *asignación aleatoria* del tratamiento como método de evaluación de impacto (para la validez interna). Una evaluación de impacto puede producir estimaciones internamente válidas del impacto mediante una asignación aleatoria del tratamiento; sin embargo, si la evaluación se lleva a cabo con una muestra no aleatoria de la población, puede que los impactos estimados no sean generalizables para el conjunto de unidades elegibles. Al contrario, si la evaluación utiliza una muestra aleatoria de la población de

unidades elegibles, pero el tratamiento no se asigna de manera aleatoria, la muestra sería representativa pero el grupo de comparación puede no ser válido, lo cual pone en entredicho la validez interna. En algunos contextos, puede que los programas se enfrenten a limitaciones que exigen un equilibrio entre validez interna y externa. Este es el caso, por ejemplo, del programa analizado anteriormente, que tiene como objetivo el 20% inferior de los hogares sobre la base del ingreso. Si este programa incorpora a todos los hogares por debajo del percentil 15º, pero lleva a cabo una evaluación de impacto de asignación aleatoria entre una muestra aleatoria de hogares entre los percentiles 15º a 25º, dicha evaluación tendrá validez interna gracias a la asignación aleatoria: es decir, se conocerá el verdadero impacto en el subconjunto de hogares entre los percentiles 15º y 25º. Sin embargo, la validez externa de la evaluación de impacto será limitada, dado que los resultados no pueden extrapolarse directamente al conjunto de la población de beneficiarios, en particular, a los hogares que se encuentren por debajo del percentil 15º.

¿Cuándo puede aplicarse la asignación aleatoria?

La asignación aleatoria puede utilizarse como regla de asignación de un programa en dos escenarios específicos:

1. *Cuando la población elegible es mayor que el número de plazas disponibles del programa.* Cuando la demanda de un programa supera a la oferta, se puede utilizar un sorteo para seleccionar el grupo de tratamiento dentro de la población elegible. En este contexto, todas las unidades de la población tienen la misma probabilidad (o una probabilidad conocida superior a 0 e inferior a 1) de ser seleccionadas para el programa. El grupo que gana el sorteo es el grupo de tratamiento y el resto de la población a la que no se ha ofrecido el programa es el grupo de comparación. Siempre que exista una limitación que impida ampliar la escala del programa a toda la población, se pueden mantener los grupos de comparación para medir los impactos del programa a corto, mediano y largo plazo. En este contexto, no hay un dilema ético en mantener indefinidamente un grupo de comparación, ya que un subgrupo de la población quedará necesariamente excluido del programa debido a problemas de capacidad.

   Por ejemplo, el ministerio de Educación desea equipar con bibliotecas a las escuelas públicas de todo el país, pero el ministerio de Finanzas solo asigna un presupuesto suficiente para cubrir una tercera parte de las bibliotecas. Si el ministerio de Educación quiere que todas las escuelas públicas tengan las mismas posibilidades de tener una biblioteca, organizará un sorteo en el que cada escuela tenga la misma probabilidad

(1 en 3) de resultar seleccionada. Las escuelas elegidas en el sorteo reciben una nueva biblioteca y constituyen el grupo de tratamiento, y a los otros dos tercios de las escuelas públicas del país no se les ofrece la biblioteca y se convierten en el grupo de comparación. A menos que se asignen más fondos al programa de bibliotecas, seguirá habiendo un grupo de escuelas que no recibirá financiamiento para una biblioteca a través del programa, y podrá usarse como grupo de comparación para medir el contrafactual.

2. *Cuando sea necesario ampliar un programa de manera progresiva hasta que cubra a toda la población elegible.* Cuando un programa se extiende por etapas, establecer de forma aleatoria el orden en el que los participantes se benefician del mismo ofrece a cada unidad elegible la misma posibilidad de recibir tratamiento en la primera fase o en una fase posterior. Siempre que no se haya sumado todavía el "último" grupo al programa, este sirve como grupo de comparación válido a partir del cual se podrá estimar el contrafactual para los que ya se han incorporado. Esta configuración también puede permitir que la evaluación recoja los efectos de una *exposición diferencial al tratamiento*, es decir, el efecto de recibir un programa durante un período más o menos prolongado.

   Por ejemplo, supóngase que el ministro de Salud quiere capacitar a los 15.000 profesionales de enfermería de todo el país en el uso de un nuevo protocolo sanitario, pero necesita tres años para capacitarlos a todos. En el contexto de una evaluación de impacto, el ministro podría seleccionar de manera aleatoria a un tercio de las enfermeras para que reciban capacitación durante el primer año, un tercio para el segundo año y un tercio para el tercer año. Para evaluar el efecto de un programa de capacitación un año después de su implementación, el grupo de enfermeras capacitadas durante el primer año constituirá el grupo de tratamiento y el grupo de enfermeras seleccionadas de modo aleatorio para recibir capacitación durante el tercer año sería el grupo de comparación, ya que todavía no se habrán expuesto al aprendizaje.

Cómo asignar aleatoriamente el tratamiento

Después de haber analizado cómo funciona la asignación aleatoria y por qué produce un buen grupo de comparación, se abordarán los pasos para asignar con éxito el tratamiento de manera aleatoria. El gráfico 4.3 ilustra este proceso.

El primer paso de la asignación aleatoria consiste en definir las unidades elegibles para el programa. Cabe recordar que, dependiendo del programa concreto, una unidad podría ser una persona, un centro de salud, una

Gráfico 4.3 Pasos para la asignación aleatoria del tratamiento

escuela, una empresa o incluso todo un pueblo o una municipalidad. La población de unidades elegibles está compuesta por aquellos para los cuales interesa conocer el impacto de un programa. Por ejemplo, si se está implementando un programa de formación para los maestros de escuela primaria en zonas rurales, los maestros de escuela primaria de zonas urbanas o los profesores de secundaria no formarían parte del conjunto de unidades elegibles.

Una vez que se ha determinado la población de unidades elegibles, habrá que comparar el tamaño del grupo con el número de observaciones requeridas para la evaluación. El tamaño de la muestra de la evaluación se establece mediante cálculos de la potencia y se basa en el tipo de preguntas a las que el evaluador desearía que se respondiera (ver capítulo 15). Si la población elegible es pequeña, quizás haya que incluir todas las unidades elegibles en la evaluación. Por el contrario, si hay más unidades elegibles de las que se requiere para la evaluación, entonces el segundo paso consiste en seleccionar una muestra de unidades a partir de la población que se incluirá en la muestra de evaluación.

Este segundo paso responde sobre todo a la necesidad de limitar los costos de la recopilación de datos. Si se observa que los datos de los sistemas

de monitoreo existentes se pueden usar para la evaluación, y que esos sistemas abarcan al conjunto de unidades elegibles, tal vez no sea necesario elaborar una muestra distinta de la evaluación. Sin embargo, imagínese una evaluación en la que la población de unidades elegibles comprende decenas de miles de maestros de todas las escuelas del país, y que se necesita recopilar información detallada sobre los conocimientos y las prácticas pedagógicas de los maestros. Entrevistar y evaluar a todos los docentes del país sería prohibitivamente oneroso e inviable en términos logísticos. A partir de los cálculos de potencia, puede que el evaluador decida que, para responder a su pregunta de interés, es suficiente contar con una muestra de 1.000 maestros distribuidos en 200 escuelas. Siempre que la muestra de docentes sea representativa del conjunto de la población de maestros, cualquier resultado de la evaluación será externamente válido y se puede generalizar al resto de los docentes del país. La recopilación de datos sobre esta muestra de 1.000 maestros en 200 escuelas será mucho menos costosa que recolectar datos sobre todos los docentes de todas las escuelas del país.

El tercer paso consiste en configurar los grupos de tratamiento y de comparación a partir de las unidades de la muestra de la evaluación, mediante la asignación aleatoria. En los casos en que la asignación aleatoria tenga que llevarse a cabo en un foro público, por ejemplo en la televisión, puede que sea necesario utilizar una técnica sencilla, como lanzar una moneda a la suerte o sacar los nombres de una caja. Los siguientes ejemplos suponen que la unidad de aleatorización es una persona individual, aunque la misma lógica se aplica a la aleatorización de más unidades agregadas de observación, como escuelas, firmas o comunidades:

1. Si se quiere asignar el 50% de los individuos al grupo de tratamiento y 50% al grupo de comparación, hay que lanzar la moneda para cada persona. Hay que decidir con antelación cuál cara de la moneda asignará una persona al grupo de tratamiento.

2. Si se quiere asignar una tercera parte de la muestra de la evaluación al grupo de tratamiento, se puede tirar un dado para cada persona. Antes, hay que decidir una regla, por ejemplo, si el dado muestra 1 o 2, el individuo será asignado al grupo de tratamiento, mientras que si arroja un 3, 4, 5 o 6 el individuo será derivado al grupo de comparación. El dado se tiraría una vez para cada persona en la muestra de evaluación, y se la asignaría sobre la base del número del dado.

3. Escribir los nombres de todos los individuos en trozos de papel de igual tamaño y forma. Plegar los papeles de modo que no se puedan leer los nombres y mezclarlos de manera conveniente en una caja o en algún otro recipiente. Antes de empezar a sacar los nombres, debe decidirse la regla,

es decir, cuántos trozos de papel se extraerán, y que extraer un nombre significa asignar a esa persona al grupo de tratamiento. Una vez que la regla esté clara, se debe solicitar a alguien del público (una persona imparcial, como un niño) que extraiga tantos trozos de papel como participantes se requiera en el grupo de tratamiento.

Si tienen que asignarse muchas unidades (por ejemplo, más de 100), utilizar enfoques sencillos como los descritos requerirá demasiado tiempo y habrá que utilizar un proceso automatizado. Para ello, primero habrá que decidir una regla de modo que se asignen los participantes sobre la base de números aleatorios. Por ejemplo, si se deben asignar 40 de 100 unidades de la muestra de evaluación al grupo de tratamiento, puede decidirse asignar esas 40 unidades con los números aleatorios más altos al grupo de tratamiento y el resto al grupo de comparación. Para implementar la asignación aleatoria, se asignará un número al azar a cada unidad en la muestra de evaluación, utilizando un generador aleatorio de números en una hoja de cálculo, o en un programa estadístico especializado (gráfico 4.4), y se utilizará la regla ya definida para formar los grupos de tratamiento y comparación. Es importante decidir la regla antes de generar los números al azar.

Gráfico 4.4 Asignación aleatoria del tratamiento mediante hoja de cálculo

	A	B	C	D	E
1	Random number	Between 0 and 1.			
2	Goal	Assign 50% of evaluation sample to treatment group.			
3	Rule	If random number is above 0.5: assign person to treatment group; otherwise: assign person to comparison group.			
5	**Unit identification**	**Name**	**Random number***	**Final random number****	**Assignment**
6	1001	Ahmed	0.7698674	0.479467635	0
7	1002	Elisa	0.4054534	0.945729597	1
8	1003	Anna	0.3584427	0.933658744	1
9	1004	Jung	0.5010306	0.383305299	0
10	1005	Tuya	0.8799600	0.102877439	0
11	1006	Nilu	0.1764322	0.228446592	0
12	1007	Roberto	0.0030776	0.444725231	0
13	1008	Priya	0.7512858	0.817004226	1
14	1009	Grace	0.1331390	0.955775449	1
15	1010	Fathia	0.8735385	0.873459852	1
16	1011	John	0.0089322	0.211028126	0
17	1012	Alex	0.0762848	0.574082414	1
18	1013	Nafula	0.5760701	0.151608805	0
19	* type the formula =RAND(). Note that the random numbers in Column C are volatile: they change everytime you do a calculation.				
20	** Copy the numbers in column C and "Paste Special>Values" into Column D. Column D then gives the final random numbers.				
21	*** type the formula =IF(C{row number}>0.5,1,0)				

De otra manera, puede que el evaluador se vea tentado de usar una regla basada en los números aleatorios que ve, lo que invalidaría la asignación aleatoria.

La lógica en que se fundamenta el proceso automatizado no es diferente de la asignación aleatoria basada en lanzar una moneda o extraer nombres de un sombrero. Se trata de un mecanismo que asigna al azar si cada unidad pertenece al grupo de tratamiento o de comparación.

Al utilizar un sorteo público, dados o números al azar generados por computador, es importante documentar el proceso para asegurar que sea transparente. En primer lugar, eso significa que la regla de asignación debe decidirse con antelación y comunicarse al público. En segundo lugar, el evaluador debe ceñirse a la regla una vez que se extraen los números al azar. En tercer lugar, debe demostrarse que el proceso era realmente aleatorio. En el caso de los sorteos y el lanzamiento de dados, se puede grabar el proceso en video; la asignación mediante números al azar generados por computador requiere que se presente un registro de los cómputos, de modo que el proceso pueda ser auditado.[5]

¿A qué nivel se lleva a cabo una asignación aleatoria?

La asignación aleatoria puede llevarse a cabo en diversos niveles: individual, hogares, empresas, comunidades o regiones. En general, el nivel en el que se asignan aleatoriamente las unidades a los grupos de tratamiento y de comparación dependerá en gran medida de dónde y cómo se implemente el programa. Por ejemplo, si se aplica un programa sanitario a nivel de las clínicas de salud, primero se elegirá una muestra aleatoria de dichas clínicas y después se asignará algunas de ellas al grupo de tratamiento y otras al grupo de comparación.

Cuando el nivel de asignación aleatoria es superior o más agregado, como el nivel regional o provincial, puede resultar difícil realizar una evaluación de impacto, porque el número de regiones o provincias en la mayoría de los países no es suficientemente grande para obtener grupos equilibrados de tratamiento y de comparación. Si un país tiene únicamente seis provincias, solo podrá haber tres de ellas en el grupo de tratamiento y tres en el grupo de comparación, lo cual es insuficiente para asegurar que las características de referencia de los grupos de tratamiento y comparación estén equilibradas. Además, para que la asignación aleatoria genere estimaciones de impacto no sesgadas, es importante garantizar que los factores externos dependientes del tiempo (como el clima o los ciclos de las elecciones locales) sean en promedio los mismos en los grupos de comparación y de tratamiento. A medida que el nivel de asignación aumenta, se vuelve cada vez más improbable que estos factores estén equilibrados entre ambos grupos.

Por ejemplo, la lluvia es un factor externo dependiente del tiempo porque varía sistemáticamente de un año al siguiente. En una evaluación del sector agrícola, convendría asegurarse de que las sequías afecten por igual a las provincias que se encuentran en el grupo de tratamiento y en el de comparación. Con solo tres provincias en los grupos de tratamiento y comparación, este equilibrio podría perderse con facilidad. Por otro lado, si se rebajara la unidad de selección al nivel subprovincial, como una municipalidad, es más probable que la lluvia esté equilibrada entre los grupos de tratamiento y comparación a lo largo del tiempo.

Asimismo, a medida que el nivel de la asignación aleatoria disminuye –por ejemplo, a nivel individual o del hogar– aumentan las probabilidades de que el grupo de comparación se vea afectado de forma involuntaria por el programa. Hay dos tipos particulares de riesgos que se deben tener en cuenta cuando se escoge el nivel de asignación, a saber: los efectos de derrame y el cumplimiento imperfecto. El *efecto de derrame* se produce cuando el grupo de tratamiento influye de forma directa o indirecta en los resultados del grupo de comparación (o viceversa). Por su parte, el *cumplimiento imperfecto* tiene lugar cuando algunos miembros del grupo de comparación participan en el programa o algunos miembros del grupo de tratamiento no lo hacen (véase un análisis más detallado de estos conceptos en el capítulo 9).

Tener en cuenta el nivel de asignación aleatoria de manera rigurosa puede minimizar el riesgo de derrame y de cumplimiento imperfecto. Las personas pueden asignarse a grupos o agrupaciones, como los alumnos de una escuela o los hogares en una comunidad, para minimizar los flujos de información y los contactos entre individuos en los grupos de tratamiento y comparación. Para reducir la contaminación, el nivel de asignación también debería escogerse según la capacidad del programa para mantener una clara diferencia entre grupos de tratamiento y comparación a lo largo de la intervención. Si el programa comprende actividades a nivel comunitario, puede que sea difícil evitar exponer a todos los individuos de esa comunidad al programa.

Un ejemplo bien conocido de efecto de derrame es la administración de medicamentos antiparasitarios a los niños. Si en el grupo de tratamiento hay hogares situados cerca de un hogar del grupo de comparación, los niños de los hogares de comparación pueden verse afectados positivamente por un efecto de derrame del grupo de tratamiento, porque se reducirán sus probabilidades de contraer parásitos procedentes de sus vecinos (Kremer y Miguel, 2004). Para aislar el impacto del programa, los hogares de tratamiento y comparación deben estar situados suficientemente lejos unos de otros de modo de evitar ese tipo de derrames. Sin embargo, a medida que la distancia entre los hogares aumente, se volverá más oneroso implementar

el programa y, a la vez, administrar las encuestas. Como regla general, si se pueden descartar los efectos de derrame de forma razonable, es preferible llevar a cabo una asignación aleatoria del tratamiento en el nivel más bajo posible de implementación del programa, lo cual garantizará que el número de unidades de los grupos de tratamiento y comparación sea el mayor posible.

La estimación del impacto bajo asignación aleatoria

Una vez que se haya seleccionado una muestra de evaluación aleatoria y asignado el tratamiento de manera aleatoria, es bastante sencillo estimar el impacto del programa. Después de que el programa ha funcionado durante un tiempo, tendrán que medirse los resultados de las unidades de tratamiento y de comparación. El impacto del programa es sencillamente la diferencia entre el resultado promedio (Y) para el grupo de tratamiento y el resultado promedio (Y) para el grupo de comparación. Por ejemplo, en el caso genérico que se presenta en el gráfico 4.5, el resultado promedio del grupo de tratamiento es 100, y el resultado promedio del grupo de comparación es 80, por lo que el impacto del programa equivale a 20. Por ahora, se supondrá que todas las unidades en el grupo de tratamiento son efectivamente tratadas y que ninguna unidad del grupo de comparación es tratada. En el ejemplo del programa de formación de los profesores, todos los profesores asignados al grupo de tratamiento reciben la formación y ninguno de los profesores del grupo de comparación la recibe. En el capítulo 5, se analiza el escenario (más realista) donde el cumplimiento es incompleto, es decir, donde menos del 100% de las unidades en el grupo de tratamiento realmente participa en la intervención o algunas unidades de comparación tienen acceso al programa. En este caso, todavía se puede obtener una estimación no sesgada del impacto del programa mediante la asignación aleatoria, aunque la interpretación de los resultados variará.

Gráfico 4.5 Estimación del impacto con la asignación aleatoria

	Tratamiento	Comparación	Impacto
	Media (Y) para el grupo de tratamiento = 100	Media (Y) para el grupo de comparación = 80	Impacto = ΔY = 20
Se inscribe si y solo si está asignado al grupo de tratamiento			

Lista de verificación: la asignación aleatoria

La asignación aleatoria es el método más robusto para estimar los contrafactuales; se considera el sello de oro de la evaluación de impacto. Para estimar la validez de esta estrategia de evaluación en un determinado contexto, deberían contemplarse algunas pruebas básicas.

- ✓ ¿Están equilibradas las características de la línea de base? Deben compararse las características de línea de base del grupo de tratamiento y del grupo de comparación.[6]
- ✓ ¿Se ha producido algún incumplimiento con la asignación? Se debe verificar si todas las unidades elegibles han recibido tratamiento y que no haya unidades no elegibles que hayan recibido tratamiento. Si ha habido incumplimiento, tendrá que utilizarse el método de variable instrumental (véase el capítulo 5).
- ✓ ¿Son suficientemente numerosas las unidades en los grupos de tratamiento y comparación? Si no, sería necesario combinar la asignación aleatoria con diferencias en diferencias (véase el capítulo 7).
- ✓ ¿Hay algún motivo para creer que los resultados en algunas unidades de alguna manera dependen de la asignación de otras unidades? ¿Podría haber un impacto del tratamiento en las unidades del grupo de comparación? (véase el capítulo 9).

Evaluación del impacto del HISP con la asignación aleatoria

Volvamos al ejemplo del Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés, *Health Insurance Subsidy Program*) y verifiquemos qué significa asignación aleatoria en este contexto. Recuérdese que se intenta estimar el impacto de un programa a partir de una prueba piloto que comprende 100 pueblos de tratamiento.

Después de llevar a cabo dos evaluaciones de impacto utilizando estimaciones potencialmente sesgadas del contrafactual en el capítulo 3 (con recomendaciones de políticas contradictorias), usted decide volver a repensar cómo obtener una estimación más precisa del contrafactual. Después de consultar con su equipo de evaluación, está convencido de que construir una estimación válida del contrafactual

requerirá identificar un grupo de pueblos que sean lo más parecidos posible a los 100 pueblos del tratamiento en todos los sentidos, con la excepción de que un grupo participó en el HISP y el otro no. Dado que el HISP se implementó como plan piloto, y que los 100 pueblos de tratamiento fueron seleccionados de forma aleatoria entre los pueblos rurales en todo el país, usted observa que los pueblos del tratamiento deberían, en promedio, tener las mismas características que los pueblos rurales no tratados en todo el país. Por lo tanto, se puede estimar el contrafactual de una manera válida, midiendo los gastos en salud de los hogares elegibles en los pueblos rurales que no participaron del programa.

Afortunadamente, en el momento de las encuestas de línea de base y de seguimiento, se recopilaron datos de otros 100 pueblos rurales a los que no se ofreció el programa. Esos 100 pueblos también fueron seleccionados de manera aleatoria entre la población de los pueblos rurales en el país. Por lo tanto, la manera en que fueron escogidos los dos grupos de pueblos garantiza que tienen características estadísticamente idénticas, excepto que los 100 pueblos de tratamiento se inscribieron en el HISP y los 100 pueblos del grupo de comparación no fueron destinatarios del programa. Se ha producido una asignación aleatoria del tratamiento.

Dada la asignación aleatoria del tratamiento, usted confía en que ningún factor externo, excepto el HISP, explicaría las diferencias en los resultados entre los pueblos de tratamiento y de comparación. Para validar este supuesto, usted comprueba si los hogares elegibles en los pueblos de tratamiento y comparación tienen características similares en la línea de base, como se muestra en el cuadro 4.1.

Usted observa que las características promedio de los hogares en los pueblos de tratamiento y de comparación son, de hecho, muy similares. Las únicas diferencias estadísticamente significativas son las relativas al número de años de escolarización del jefe de hogar y la distancia al hospital, y esas diferencias son pequeñas (solo 0,16 años, o menos del 6% de los años de escolarización promedio del grupo de comparación, y 2,91 km, o menos del 3% de la distancia promedio al hospital del grupo de comparación). Incluso con un experimento aleatorio en una muestra grande, se puede esperar un pequeño número de diferencias debido al azar y a las propiedades del test estadístico. De hecho, al utilizar niveles de significancia estándar del 5%, podía esperarse que alrededor del 5% de las diferencias en las características sean estadísticamente significativas, aunque no se esperaría que la magnitud de estas diferencias fuese grande.

Cuadro 4.1 Evaluación del HISP: balance entre los pueblos de tratamiento y de comparación en la línea de base

Características de los hogares	Pueblos de tratamiento (N = 2964)	Pueblos de comparación (N = 2664)	Diferencia	*t*-estadístico
Gasto en salud (dólares de EE.UU. anuales per cápita)	14,49	14,57	–0,08	–0,73
Edad del jefe de hogar (años)	41,66	42,29	–0,64	–1,69
Edad del cónyuge (años)	36,84	36,88	0,04	0,12
Nivel de estudios del jefe de hogar (años)	2,97	2,81	0,16*	2,30
Nivel de estudios del cónyuge (años)	2,70	2,67	0,03	0,43
Jefe de hogar es mujer = 1	0,07	0,08	–0,01	–0,58
Jefe de hogar es indígena = 1	0,43	0,42	0,01	0,69
Número de miembros del hogar	5,77	5,71	0,06	1,12
Tiene suelo de tierra	0,72	0,73	–0,01	–1,09
Tiene baño = 1	0,57	0,56	0,01	1,04
Hectáreas de terreno	1,68	1,72	–0,04	–0,57
Distancia a un hospital (km)	109,20	106,29	2,91	2,57

** Significativo al nivel del 1%.

Con la validez del grupo de comparación ya establecida, ahora se puede estimar el contrafactual como los gastos promedio en salud de los hogares elegibles en los 100 pueblos del grupo de comparación. El cuadro 4.2 muestra los gastos promedio en salud de los hogares elegibles en los pueblos de los grupos de tratamiento y de comparación. Nótese que en la línea de base los gastos promedio en salud de los

Cuadro 4.2 Evaluación del HISP según la asignación aleatoria (comparación de medias)

	Tratamiento	Comparación	Diferencia	*t*-estadístico
Línea de base: gasto en salud de los hogares (en dólares de EE.UU.)	14,49	14,57	−0,08	−0,73
Encuesta de seguimiento: gasto en salud de los hogares (en dólares de EE.UU.)	7,84	17,98	−10,14**	−49,15

** Significativo al nivel del 1%.

hogares de los grupos de tratamiento y comparación no son estadísticamente diferentes, como debería esperarse con una asignación aleatoria.

Ahora que se cuenta con un grupo de comparación válido, se puede encontrar el impacto del HISP sencillamente calculando la diferencia entre los gastos directos promedio en salud de los hogares en los pueblos de tratamiento y de comparación asignados de forma aleatoria en el período de seguimiento. El impacto es una reducción de US$10,14 a lo largo de dos años. Replicar este resultado mediante un análisis de regresión lineal arroja el mismo resultado, como se observa en el cuadro 4.3. Por último, mediante un análisis de regresión multivariante que controla por otras características observables de los hogares de la muestra, se observa que el programa ha reducido los gastos de los hogares inscritos en US$10,01, a lo largo de dos años, lo cual es casi idéntico al resultado de la regresión lineal.

Con la asignación aleatoria, podemos estar seguros de que no hay factores que sean sistemáticamente diferentes entre los grupos de tratamiento y comparación que también puedan explicar la diferencia en gastos en salud. Ambos conjuntos de pueblos comenzaron con características promedio muy similares y han estado expuestos al mismo conjunto de políticas y programas nacionales durante los dos años de tratamiento. Por lo tanto, el único motivo plausible por el que los hogares pobres en las comunidades de tratamiento tienen gastos inferiores a los de los hogares

en los pueblos de comparación es que el primer grupo fue destinatario del programa de seguro de salud y el otro grupo no lo fue.

Cuadro 4.3 Evaluación del HISP según la asignación aleatoria (análisis de regresión)

	Regresión lineal	Regresión lineal multivariante
Impacto estimado sobre el gasto en salud de los hogares	−10,14**	−10,01**
	(0,39)	(0,34)

Nota: Los errores estándares están entre paréntesis.

** Significativo al nivel del 1%.

Pregunta HISP 3

A. ¿Por qué la estimación de impacto obtenida mediante una regresión lineal multivariante se mantiene básicamente constante cuando se controla por otros factores, al cotejarse con la regresión lineal simple y la comparación de medias?

B. Sobre la base del impacto estimado con el método de asignación aleatoria, ¿debería ampliarse el HISP a nivel nacional?

Recursos adicionales

- Para material de apoyo para este capítulo e hipervínculos de recursos adicionales, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para más recursos sobre las evaluaciones de impacto con asignación aleatoria, véase el portal de evaluación del BID (www.iadb.org/portalevaluacion).
- Para un resumen completo de las evaluaciones de impacto con asignación aleatoria, véase el siguiente libro y el sitio web correspondiente:
  - ➢ R. Glennerster y K. Takavarasha (2013), *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press (http://runningres.com/).

- Para un debate en profundidad sobre cómo encontrar el equilibrio entre grupos de tratamiento y de comparación mediante la asignación aleatoria, véase:
  - ➢ M. Bruhn y D. McKenzie (2009), "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1(4): 200–32.
- Para un ejemplo de asignación aleatoria pública para una evaluación en Camerún, véase el World Bank Impact Evaluation Toolkit, Módulo 3 (www.worldbank.org/health/impactevaluationtoolkit).

Notas

1. La asignación aleatoria del tratamiento también suele denominarse *ensayo aleatorio controlado*, *evaluaciones aleatorias*, *evaluaciones experimentales* y *experimentos sociales*, entre otras denominaciones. Estrictamente hablando, un experimento no tiene que identificar impactos mediante asignaciones aleatorias, pero los evaluadores suelen utilizar el término "experimento" solo cuando la evaluación recurre a la asignación aleatoria.
2. Nótese que esta probabilidad no necesariamente significa una probabilidad del 50% de ganar el sorteo. En la práctica, la mayoría de las evaluaciones con asignación aleatoria darán a cada unidad elegible una probabilidad de selección determinada, de manera que el número de ganadores (tratamientos) sea igual al total de beneficios disponibles. Por ejemplo, si un programa tiene suficientes fondos para servir solo a 1.000 comunidades de una población de 10.000 comunidades elegibles, cada comunidad tendrá una probabilidad de una entre 10 de ser seleccionada para el tratamiento. La potencia estadística (un concepto analizado más en detalle en el capítulo 15) se maximizará cuando la muestra de evaluación se divida por igual entre los grupos de tratamiento y comparación. En el ejemplo de este caso, para un tamaño total de la muestra de 2.000 comunidades, la potencia estadística se maximizará si se seleccionan las 1.000 comunidades de tratamiento y una submuestra de 1.000 comunidades de comparación, en lugar de tomar una muestra aleatoria simple del 20% de las 10.000 comunidades originales elegibles (lo que produciría una muestra de evaluación de alrededor de 200 comunidades de tratamiento y 1.800 comunidades de comparación).
3. Por ejemplo, los programas de vivienda que otorgan viviendas subvencionadas suelen utilizar los sorteos para seleccionar a los participantes del programa. Numerosas escuelas subvencionadas en Estados Unidos seleccionan a los postulantes mediante sorteo.
4. Además de crear grupos que tienen características promedio similares, la asignación aleatoria también crea grupos que tienen distribuciones similares.
5. La mayoría de los programas informáticos permiten establecer un número aleatorio para que los resultados de la asignación aleatoria sean plenamente transparentes y replicables.

6. Como se ha mencionado, por motivos estadísticos no todas las características observables deben ser similares en los grupos de tratamiento y de comparación para que la aleatorización sea exitosa. Incluso cuando las características de los dos grupos son verdaderamente idénticas, se puede esperar que el 5% de las mismas aparecerán con una diferencia estadísticamente significativa cuando se utiliza un intervalo de confianza de 95% para la prueba. Las variables en cuyo caso se presenta una diferencia grande entre los grupos de tratamiento y de comparación son especialmente preocupantes.

Referencias bibliográficas

Bertrand, M., B. Crépon, A. Marguerie y P. Premand. 2016. "Impacts à Court et Moyen Terme sur les Jeunes des Travaux à Haute Intensité de Main d'oeuvre (THIMO): Résultats de l'évaluation d'impact de la composante THIMO du Projet Emploi Jeunes et Développement des Compétence (PEJEDEC) en Côte d'Ivoire." Washington, D.C.: Banco Mundial y Abidjan, BCP-Emploi.

Blattman, C., N. Fiala y S. Martínez. 2014. "Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda." *Quarterly Journal of Economics*. (doi: 10.1093/qje/qjt057.).

Bruhn, M. y D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1(4): 200–32.

Dupas, P. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3 (1): 1–34.

Glennerster, R. y K. Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press.

Kremer, M., J. Leino, E. Miguel y A. Peterson Zwane. 2011. "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions." *Quarterly Journal of Economics* 126: 145–205.

Kremer, M. y E. Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

Premand, P., O. Barry y M. Smitz. 2016. "Transferts monétaires, valeur ajoutée de mesures d'accompagnement comportemental, et développement de la petite enfance au Niger. Rapport descriptif de l'évaluation d'impact à court terme du Projet Filets Sociaux." Washington, D.C.: Banco Mundial.

Schultz, P. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.

CAPÍTULO 5

Las variables instrumentales

La evaluación de programas cuando no todos cumplen su asignación

En el análisis de la asignación aleatoria del capítulo 4, se asume que el administrador del programa tiene la facultad para asignar la intervención a los grupos de tratamiento y de comparación, y que los asignados al tratamiento participan en el programa y los asignados al grupo de comparación no lo hacen. En otras palabras, las observaciones asignadas a los grupos de tratamiento y de comparación cumplen su asignación. El pleno cumplimiento se logra con mayor frecuencia en pruebas de laboratorio o en ensayos médicos, donde el investigador puede asegurar, primero, que todos los sujetos del grupo de tratamiento reciban un determinado tratamiento y, segundo, que no lo reciba ninguno de los sujetos del grupo de comparación.[1] De manera más general, en el capítulo 4 se asume que los programas pueden determinar quiénes son los participantes potenciales, excluyendo a algunos y asegurando que otros participen.

Sin embargo, en los programas sociales del mundo real, puede que sea poco realista pensar que el administrador del programa será capaz de asegurar un cumplimiento pleno de la asignación del grupo. Aun así, numerosos programas permiten a los participantes potenciales elegir si se inscriben o no y, por lo tanto, no pueden excluir a participantes potenciales que quieran inscribirse. Además, algunos programas tienen un presupuesto lo suficientemente grande para administrar la intervención de forma inmediata

Concepto clave

El método de variables instrumentales se sustenta en alguna fuente externa de variación para determinar el estatus del tratamiento. Ejerce una influencia sobre la probabilidad de participar en un programa, pero está fuera del control de los participantes y no se relaciona con las características de los mismos.

a toda la población elegible, de modo que no sería ético asignar aleatoriamente a las personas a los grupos de tratamiento y de comparación, y excluir a participantes potenciales en aras de la evaluación. Por ende, se necesita una manera alternativa de evaluar el impacto de este tipo de programas.

El método denominado *variables instrumentales* (VI) puede resultar útil para evaluar los programas con cumplimiento imperfecto, inscripción voluntaria o cobertura universal. En general, para estimar los impactos, el método VI depende de una fuente externa de variación para determinar el estatus del tratamiento. El método puede aplicarse a un amplio espectro de situaciones, más allá de la evaluación de impacto. Se puede pensar en una VI como algo que escapa al control del individuo y que influye en su probabilidad de participar en un programa pero que, de otra manera, no está asociado con las características de dicho individuo.

En este capítulo, se analizará cómo esta variación externa, o VI, puede ser generada por las reglas de funcionamiento del programa que está bajo el control de los encargados del mismo o de los equipos de evaluación. Para producir evaluaciones de impacto válidas, esta fuente externa de variación, o VI, debe satisfacer un cierto número de condiciones, las cuales se abordarán detenidamente en este capítulo. Se ha observado que la asignación aleatoria del tratamiento, como se señaló en el capítulo 4, es un muy buen instrumento y que satisface las condiciones necesarias. El método VI se utilizará en dos aplicaciones comunes de la evaluación de impacto. Primero, se usará como una extensión del método de asignación aleatoria cuando no todas las unidades cumplen con su asignación de grupo. En segundo lugar, se recurrirá a él para diseñar una promoción aleatoria del tratamiento, un método de evaluación que puede funcionar en algunos programas que ofrecen inscripción voluntaria o cobertura universal. El recuadro 5.1 ilustra un uso creativo del método de VI.

Tipos de estimaciones de impacto

Una evaluación de impacto siempre estima el impacto de un programa comparando los resultados de un grupo de tratamiento con la estimación de un contrafactual obtenido de un grupo de comparación válido. En el capítulo 4 se asumía que había *pleno cumplimiento* en el tratamiento, es decir, que todas las unidades a las que se ofrecía un programa se inscribían en él y que ninguna de las unidades del grupo de comparación recibía el programa. En este escenario, se estimaba el efecto promedio del tratamiento para la población.

En la evaluación de los programas en el mundo real, donde los participantes potenciales pueden decidir si se inscriben o no, el pleno

Recuadro 5.1: El uso de variables instrumentales para evaluar el impacto de *Plaza Sésamo* en la preparación escolar

El programa de televisión *Plaza Sésamo*, destinado a preparar a los niños en edad preescolar para la escuela primaria, obtuvo rápidamente la aclamación de la crítica y gozó de gran popularidad al emitirse por primera vez en 1969. Desde entonces lo han visto millones de niños. En 2015 Kearney y Levine se propusieron estudiar los impactos a largo plazo del programa en una evaluación retrospectiva realizada en Estados Unidos. Aprovechando las limitaciones de la tecnología de las emisiones televisivas durante los primeros años del programa, los investigadores utilizaron un enfoque de variables instrumentales (VI).

En los primeros años, el programa no llegaba a todos los hogares. Solo se emitía en los canales de frecuencia ultra alta (UHF). Así, alrededor de solo dos tercios de la población de Estados Unidos vivía en zonas con acceso al programa. Por lo tanto, Kearney y Levine (2015) utilizaron la distancia entre los hogares y la antena de televisión más cercana que transmitiera UHF como instrumento de participación en el programa. Los investigadores sostenían que, dado que las antenas de televisión estaban construidas en lugares escogidos por el gobierno –antes de que *Plaza Sésamo* comenzara a ser emitido– la variable no estaría relacionada con las características del hogar ni con cambios en el resultado.

La evaluación encontró resultados positivos en la preparación escolar de los niños en edad preescolar. En las zonas con recepción de la señal de televisión UHF cuando el programa comenzó, los niños tenían más probabilidades de cursar la escuela primaria a la edad adecuada. Este efecto fue notable en los niños afroamericanos y en los niños no hispanos, en los varones y en los pequeños de zonas económicamente desfavorecidas.

Fuente: Kearney y Levine (2015).

cumplimiento es menos común que en contextos como los experimentos de laboratorio. En la práctica, los programas suelen ofrecer tratamiento a un grupo específico, y algunas unidades participan y otras no. En este caso, sin pleno cumplimiento, las evaluaciones de impacto pueden estimar el efecto de *ofrecer* un programa o el efecto de *participar* en el programa.

La *intención de tratar* (ITT, por sus siglas en inglés, *intention-to-treat*) es un promedio ponderado de los resultados de los participantes y no participantes en el grupo de tratamiento versus el resultado promedio del grupo de comparación. Es importante en aquellos casos en los que se intenta determinar el impacto promedio de ofrecer un programa y la inscripción en el grupo de tratamiento es voluntaria. En cambio, puede que también se desee conocer el impacto de un programa en el grupo de individuos a los que se ofrece el programa y que realmente participan.

Concepto clave

La intención de tratar (ITT) estima la diferencia en los resultados entre las unidades asignadas al grupo de tratamiento y las unidades asignadas al grupo de comparación, independientemente de si las unidades asignadas al grupo de tratamiento reciben en efecto el tratamiento.

Concepto clave

El tratamiento en los tratados (TOT) estima la diferencia en los resultados entre las unidades que realmente reciben el tratamiento y el grupo de comparación.

Este impacto estimado se denomina *tratamiento en los tratados* (TOT, por sus siglas en inglés, *treatment-on-the-treated*). La ITT y el TOT serán iguales cuando haya pleno cumplimiento. Más adelante se volverá sobre las diferencias entre ambos, pero ahora se presentará un ejemplo para ilustrar estos conceptos.

Considérese el Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés), que se ha analizado en los capítulos anteriores. Debido a consideraciones operativas y para minimizar los efectos de derrame, la unidad de asignación del tratamiento elegida por el gobierno es el centro poblado. Los hogares de un centro poblado de tratamiento (las circunscripciones donde se ofrece el programa de seguro de salud) pueden inscribirse de forma voluntaria para un subsidio de seguro de salud, mientras que los hogares de las comunidades de comparación no pueden. A pesar de que todos los hogares de los pueblos de tratamiento son elegibles para inscribirse en el programa de seguro de salud, puede que una parte de los mismos –por ejemplo, el 10%– decida no hacerlo (quizá porque ya tienen un seguro a través de sus empleos, porque están sanos y no prevén la necesidad de cuidados sanitarios, o por muchos otros motivos).

En este escenario, el 90% de los hogares del pueblo de tratamiento decide inscribirse en el programa y recibe los servicios que este ofrece. La estimación de ITT se obtendría comparando el resultado promedio de todos los hogares a los que se ofreció el programa –es decir, el 100% de los hogares de los centros poblados de tratamiento– con el resultado promedio en los pueblos de comparación (donde no se ha inscrito ningún hogar). En cambio, el TOT se puede ver como el impacto estimado para el 90% de los hogares de las circunscripciones de tratamiento que se inscribieron en el programa. Cabe señalar que el impacto del TOT no es necesariamente el mismo que el impacto que se obtendría para el 10% de los hogares de los pueblos de tratamiento que no se inscribieron, en caso de que se inscriban. Esto es debido a que los individuos que participan en un programa cuando el mismo es ofrecido pueden ser distintos a los individuos a los que se ofrece el programa pero que deciden no inscribirse. Por ello, los efectos del tratamiento "local" no pueden extrapolarse directamente de un grupo a otro.

El cumplimiento imperfecto

Como ya se ha señalado, en los programas sociales del mundo real, el pleno cumplimiento con los criterios de selección de un programa (y, por ende, la adhesión a la condición de tratamiento o comparación) es deseable, y los responsables de las políticas y los equipos de evaluación por igual suelen intentar acercarse lo más posible a ese ideal. Sin embargo, en la práctica, no

siempre se consigue un cumplimiento del 100% de las asignaciones a los grupos de tratamiento y comparación, a pesar de los esfuerzos del encargado del programa y del equipo de evaluación. A continuación, se presentarán diferentes casos que pueden ocurrir y se debatirán las implicaciones para los métodos de evaluación que se pueden utilizar. Para empezar, hay que subrayar que la mejor solución para el cumplimiento imperfecto consiste sencillamente en evitarlo. En este sentido, los administradores del programa y los responsables de las políticas deberían intentar que el cumplimiento sea lo más alto posible en el grupo de tratamiento y lo más bajo posible en el grupo de comparación.

Supóngase que se intenta evaluar un programa de formación docente, para el cual son elegibles 2.000 maestros a fin de que participen en una capacitación piloto. Los maestros han sido asignados de forma aleatoria a uno de dos grupos: 1.000 al de tratamiento y otros 1.000 al de comparación. Cuando todos los docentes del grupo de tratamiento reciben la capacitación, y ninguno en el grupo de comparación la ha recibido, se estima el efecto promedio del tratamiento (ATE, por sus siglas en inglés, *average treatment effect*) calculando la diferencia en los resultados medios (por ejemplo, las puntuaciones en las pruebas de los alumnos) entre los dos grupos. Este ATE es el impacto promedio del tratamiento en los 1.000 maestros, dado que todos los maestros asignados al grupo de tratamiento realmente asisten al curso, algo que no ocurre con ninguno de los maestros asignados al grupo de comparación.

El primer caso de cumplimiento imperfecto ocurre cuando algunas unidades asignadas al grupo de tratamiento deciden no inscribirse o, por algún otro motivo, no reciben tratamiento. En el ejemplo de la formación docente, algunos maestros asignados al grupo de tratamiento no se presentan el primer día del curso. En este caso, no se puede calcular el tratamiento promedio para todos los maestros porque algunos nunca se inscribieron; por lo tanto, nunca se podrá calcular qué resultados habrían tenido con el tratamiento. Sin embargo, se puede estimar el impacto promedio del programa en aquellos que realmente siguen o aceptan el tratamiento. Se quiere estimar el impacto del programa en aquellos maestros a los que se asignó el tratamiento y que en la práctica se inscribieron. Esta es la *estimación del TOT*. En el ejemplo de la formación docente, la estimación del TOT representa el impacto en los maestros asignados al grupo de tratamiento que se presentaron y recibieron la capacitación.

El segundo caso de cumplimiento imperfecto se produce cuando los individuos asignados al grupo de comparación consiguen participar en el programa. En este caso, los impactos no pueden estimarse directamente para todo el grupo de tratamiento porque sus "contrapartes" en el grupo de comparación no se pueden observar sin tratamiento. Se suponía que las

unidades tratadas en el grupo de comparación generaban una estimación del contrafactual para algunas unidades en el grupo de tratamiento, pero en la práctica reciben el tratamiento; por lo tanto, no hay manera de saber cuál habría sido el impacto del programa en este subconjunto de individuos. En el ejemplo de la formación docente, supóngase que los maestros más motivados del grupo de comparación consiguen asistir de alguna manera al curso. En este caso, los más motivados en el grupo de tratamiento no tendrían contrapartes en el grupo de comparación, de modo que no sería posible estimar el impacto de la formación en ese segmento de maestros motivados.

Cuando hay incumplimiento en cualquiera de los dos lados, debería pensarse detenidamente en qué tipo de efecto de tratamiento se estima y cómo interpretarlo. Una primera opción consiste en calcular una comparación del grupo originalmente asignado al tratamiento con el grupo originalmente asignado a la comparación; esto dará la *estimación de la ITT*. La ITT compara a aquellos a quienes se pretende tratar (los asignados al grupo de tratamiento) con aquellos que se intenta no tratar (los asignados al grupo de comparación). Si el incumplimiento se produce solo del lado del tratamiento, puede ser una medida de impacto interesante y relevante, porque en cualquier caso la mayoría de los responsables de las políticas y administradores de programa solo pueden ofrecer un programa y no pueden obligar a su población designada a aceptar el mismo.

En el ejemplo de la formación docente, puede que el gobierno quiera conocer el impacto promedio del programa en todos los maestros asignados, aunque algunos de ellos no asistan al curso. Esto se debe a que, aunque el gobierno amplíe el programa, es probable que haya maestros que nunca asistirán. Sin embargo, si hay incumplimiento en el lado de la comparación, la estimación de la ITT no es tan esclarecedora. En el caso de la formación docente, dado que el grupo de comparación incluía a maestros formados, el resultado promedio en el grupo de comparación se ha visto afectado por el tratamiento. Supóngase que el efecto de la formación docente en los resultados es positivo. Si aquellos que incumplieron en el grupo de comparación son los maestros más motivados y los que más se benefician de la capacitación, el resultado promedio para el grupo de comparación tendrá un sesgo positivo (porque los maestros motivados del grupo de comparación que recibieron capacitación harán subir el resultado promedio) y la estimación ITT tendrá un sesgo negativo (dado que se trata de la diferencia entre los resultados promedio en el grupo de tratamiento y de comparación).

En estas circunstancias de no cumplimiento, una segunda opción consiste en estimar lo que se conoce como el *efecto local promedio del tratamiento* (LATE, por sus siglas en inglés, *local average treatment effect*). El LATE debe ser interpretado con cuidado, ya que representa los efectos

del programa solo para un subgrupo específico de la población. En particular, cuando hay incumplimiento en el grupo de tratamiento y en el de comparación, el LATE es el impacto en el subgrupo de cumplidores. En el ejemplo de la formación docente, si hay incumplimiento en ambos grupos, la estimación LATE es válida solo para los maestros del grupo de tratamiento que se inscribieron en el programa y que no se habrían inscrito si hubieran sido asignados al grupo de comparación.

A continuación, se explicará cómo estimar el LATE y, algo que es igual de importante, cómo interpretar los resultados. Los principios para estimar el LATE se aplican cuando hay incumplimiento en el grupo de tratamiento, en el de comparación, o en ambos al mismo tiempo. El TOT es simplemente un LATE en el caso más específico en que hay incumplimiento solamente en el grupo de tratamiento. Por lo tanto, el resto de este capítulo se enfoca en cómo estimar el LATE.

Asignación aleatoria de un programa y aceptación final

Imagínese que se debe evaluar el impacto de un programa de formación para el empleo en los salarios de los individuos. El programa se asigna de forma aleatoria a nivel individual. El grupo de tratamiento recibe el programa, mientras que el grupo de comparación no lo recibe. Lo más probable es que se encuentren tres tipos de individuos en la población:

- *Inscritos si se lo ofrecen.* Son los individuos que cumplen con su asignación. Si se les asigna al grupo de tratamiento (asignados al programa), lo aceptarán y se inscribirán. Si se les asigna al grupo de comparación (no asignados al programa), no se inscriben.
- *Nuncas.* Son los individuos que jamás se inscriben ni aceptan el programa, aunque se les asigne al grupo de tratamiento. Si en efecto se les asigna a este último, serán incumplidores.
- *Siempres.* Estos individuos encontrarán una manera de inscribirse en el programa o aceptarlo, aunque se les asigne al grupo de comparación. Si en efecto son asignados al grupo de comparación, serán incumplidores.

En el contexto de un programa de formación laboral, el grupo de los *Nuncas* puede estar formado por personas no motivadas que, aunque se les asigne un lugar en el curso, no se presentan. En cambio, los del grupo de los *Siempres* están tan motivados que encuentran una manera de entrar en el programa aunque originalmente se les haya asignado al grupo de comparación. El grupo de *Inscritos si se lo ofrecen* son los que se anotan en el curso si son asignados a él, pero no buscan inscribirse si son asignados al grupo de comparación.

El gráfico 5.1 presenta la asignación aleatoria del programa y de la inscripción final, o aceptación, cuando los tipos *Inscritos si se lo ofrecen*, *Nuncas* y *Siempres* están presentes. Supóngase que la población se compone de un 80% de *Inscritos si se lo ofrecen*, 10% de *Nuncas* y 10% de *Siempres*. Si se toma una muestra aleatoria de la población para la muestra de evaluación, dicha muestra tendrá también alrededor de un 80% de *Inscritos si se lo ofrecen*, 10% de *Nuncas* y 10% de *Siempres*. Luego, si la muestra de evaluación se asigna de manera aleatoria a un grupo de tratamiento y a un grupo de comparación, una vez más debería haber cerca de un 80% de *Inscritos si se lo ofrecen*, 10% de *Nuncas* y 10% de *Siempres* en ambos grupos. En el grupo asignado al tratamiento, se inscribirán los individuos *Inscritos si se lo ofrecen* y *Siempres*, y solo el grupo de *Nuncas* permanecerá al margen. En el grupo de comparación, los individuos de *Siempres* se inscribirán, mientras que los grupos de *Inscritos si se lo ofrecen* y *Nuncas* permanecerán fuera. Es importante recordar que si bien se sabe que en la población existen estos tres tipos de individuos, no es posible necesariamente distinguir el tipo de un individuo hasta que se observan ciertas conductas. En el grupo de tratamiento, se podrá identificar a los tipos de *Nuncas* cuando no se inscriben, pero no se podrá distinguir entre los *Inscritos si se lo ofrecen* y los *Siempres*, dado que ambos tipos se inscribirán. En el grupo de comparación, se podrá identificar a

Gráfico 5.1 Asignación aleatoria con cumplimiento imperfecto

los *Siempres* cuando se inscriben, pero no se podrá distinguir entre los *Inscritos si se lo ofrecen* y los *Nuncas*, dado que ninguno de los dos tipos se inscribirá.

La estimación de impacto bajo asignación aleatoria con cumplimiento imperfecto

Después de establecer la diferencia entre asignar un programa y la inscripción o aceptación en la práctica, se estimará el LATE del programa. Esta estimación se lleva a cabo en dos pasos, los cuales se ilustran en el gráfico 5.2.[2]

Para estimar los impactos del programa bajo la asignación aleatoria con cumplimiento imperfecto, primero se estima el impacto de la ITT. Se debe recordar que se trata solo de la diferencia en el indicador de resultados (Y)

Gráfico 5.2 Estimación del efecto local promedio del tratamiento bajo asignación aleatoria con cumplimiento imperfecto

	Grupo asignado al tratamiento	Grupo no asignado al tratamiento	Impacto
	Porcentaje inscrito = 90% Media Y para los asignados a tratamiento = 110	Porcentaje inscrito = 10% Media Y para los no asignados a tratamiento = 70	Δ porcentaje de inscritos = 80% ΔY = ITT = 40 LATE = 40%/80% = 50
Nunca se inscribe			—
Solo se inscribe si es asignado			
Siempre se inscribe			—

Nota: La estimación de la intención de tratar (ITT) se obtiene comparando los resultados de los individuos asignados al grupo de tratamiento con los de aquellos asignados al grupo de comparación, independientemente de la inscripción en la práctica. La estimación del efecto local promedio del tratamiento (LATE) es el impacto del programa en los que se inscriben solo si son asignados al programa (*Inscritos si se lo ofrecen*). La estimación LATE no proporciona el impacto del programa en aquellos que nunca se inscriben (*Nuncas*) o en aquellos que siempre se inscriben (*Siempres*).
Δ = impacto causal; Y = resultado.

para el grupo que se asigna al tratamiento y el mismo indicador para el grupo al que no se asigna tratamiento. Por ejemplo, si el salario medio (*Y*) del grupo de tratamiento es US$110 y el salario medio del grupo de comparación es US$70, la estimación de la ITT del impacto sería de US$40 (US$110 menos US$70).

En segundo lugar, habría que recuperar la estimación del LATE para el grupo de *Inscritos si se lo ofrecen* de la estimación ITT. Para esto, se debe identificar de dónde proviene la diferencia de US$40. Se procede por eliminación. Primero, se sabe que la diferencia no puede ser causada por diferencias entre las personas que nunca se inscriben (los *Nuncas*) en los grupos de tratamiento y de comparación. Esto se debe a que los *Nuncas* jamás se inscriben en el programa, de modo que para ellos es igual estar en el grupo de tratamiento que en el de comparación. En segundo lugar, se sabe que la diferencia de US$40 no puede ser producida por diferencias entre los individuos de *Siempres* en los grupos de tratamiento y de comparación porque estos siempre se anotan en el programa. Para ellos tampoco hay diferencia entre estar en el grupo de tratamiento o el grupo de comparación. Por lo tanto, la diferencia en los resultados entre ambos grupos debe necesariamente provenir del efecto del programa en el único grupo afectado por su asignación al grupo de tratamiento o de comparación, es decir, el grupo de *Inscritos si se lo ofrecen*. Si se puede identificar a los *Inscritos si se lo ofrecen* en ambos grupos, será fácil estimar el impacto del programa en ellos.

En realidad, aunque se sabe que estos tres tipos de individuos existen en la población, no se puede separar a los individuos en función de si son *Inscritos si se lo ofrecen*, *Nuncas* o *Siempres*. En el grupo que fue asignado al tratamiento, se puede identificar a los *Nuncas* (porque no se han inscrito), pero no se puede diferenciar entre los *Siempres* y los *Inscritos si se lo ofrecen* (porque ambos están inscritos). En el grupo de comparación, se puede identificar el grupo de *Siempres* (porque se inscriben en el programa), pero no es posible diferenciar entre los *Nuncas* y los *Inscritos si se lo ofrecen*.

Sin embargo, una vez que se observa que el 90% de las unidades en el grupo asignado al tratamiento en efecto se inscribe, se puede deducir que el 10% de las unidades de nuestra población debe estar formada *Nuncas* (es decir, el porcentaje de individuos del grupo asignados al tratamiento que no se inscribieron). Además, si se observa que el 10% de las unidades del grupo de comparación se inscribe, se sabe que el 10% son *Siempres* (una vez más, el porcentaje de individuos de nuestro grupo que no fue asignado al tratamiento y que sí se inscribió). Esto deja al 80% de las unidades en el grupo de *Inscritos si se lo ofrecen*. Se sabe que el impacto de US$40 proviene de una diferencia en la inscripción en el 80% de las unidades de la muestra que corresponde a *Inscritos si se lo ofrecen*. Si el 80% de las unidades es

responsable de un impacto promedio de US$40 en el conjunto del grupo asignado al tratamiento, el impacto en ese 80% de *Inscritos si se lo ofrecen* debe ser 40/0,8, o US$50. Dicho de otra manera, el impacto del programa para los *Inscritos si se lo ofrecen* es de US$50, pero cuando este impacto se distribuye en el conjunto del grupo asignado al tratamiento, el efecto promedio se diluye debido al 20% que no cumplió con la asignación aleatoria original.

Recuérdese que uno de los problemas básicos de la autoselección en los programas es que no siempre se puede saber por qué algunas personas deciden participar y otras no. Cuando se lleva a cabo una evaluación donde las unidades están asignadas de forma aleatoria, pero la participación en la práctica es voluntaria o existe una forma en que las unidades del grupo de comparación participen en el programa, se presenta un problema similar, a saber, que no siempre se entenderá la conducta que determina si un individuo se comporta como un *Nunca*, un *Siempre*, o un *Inscrito si se lo ofrecen*. Sin embargo, si la falta de cumplimiento no es demasiado severo, la asignación aleatoria sigue proporcionando un instrumento útil para la evaluación del impacto. El aspecto negativo de la asignación aleatoria con cumplimiento imperfecto es que esta estimación de impacto ya no es válida para el conjunto de la población. En cambio, la estimación debería interpretarse como una estimación local que se aplica solo a un subgrupo específico dentro de la población designada, los *Inscritos si se lo ofrecen*.

La asignación aleatoria de un programa tiene dos características importantes que permiten estimar el impacto cuando hay cumplimiento imperfecto (véase el recuadro 5.2):

1. Puede servir para predecir la inscripción en el programa si la mayoría de las personas se comportan como *Inscritos si se lo ofrecen*, y se inscriben en el programa cuando se les asigna al tratamiento y no se inscriben cuando no se les asigna.

2. Dado que los dos grupos (asignados y no asignados al tratamiento) se generan mediante un proceso aleatorio, las características de los individuos en los dos grupos no están correlacionadas con ningún otro factor –como la habilidad o la motivación– que pueda influir también en los resultados (Y).

En términos estadísticos, la asignación aleatoria sirve como VI. Se trata de una variable que predice la inscripción real de unidades en un programa, pero que no está relacionada con otras características de los individuos que puedan estar vinculadas a los resultados. Aunque en parte la decisión de los individuos de inscribirse en un programa no puede estar controlada por los administradores del programa, otra parte de la decisión sí está bajo

Recuadro 5.2: Variables instrumentales para lidiar con la falta de cumplimiento en un programa de vales escolares en Colombia

El Programa de Ampliación de Cobertura de la Educación Secundaria (PACES), en Colombia, suministró vales a más de 125.000 estudiantes para cubrir algo más de la mitad del costo de asistencia a escuelas secundarias privadas. Dado el presupuesto limitado del programa, los vales se asignaron mediante sorteo. Angrist et al. (2002) aprovecharon este tratamiento asignado de manera aleatoria para determinar el efecto del programa sobre los resultados educativos y sociales.

Angrist et al. (2002) observaron que los ganadores del sorteo tenían un 10% más de probabilidades de terminar el octavo grado y registraron una desviación estándar de 0,2 puntos más en las pruebas estandarizadas tres años después del sorteo inicial. También observaron que los efectos educativos fueron mayores en las niñas que los niños. Luego examinaron el impacto del programa en varios resultados no educativos y observaron que era menos probable que los ganadores del sorteo estuvieran casados y que trabajaban alrededor de 1,2 horas menos por semana.

Hubo cierto incumplimiento con el diseño aleatorio, ya que alrededor del 90% de los ganadores del sorteo habían usado el vale u otra forma de beca, y el 24% de los perdedores del sorteo habían recibido becas. Utilizando nuestra terminología, la población debe haber contenido un 10% de *Nuncas*, un 24% de *Siempres* y un 66% de *Inscritos si se lo ofrecen*. Angrist et al. (2002) también utilizaron la asignación original, o la condición del ganador o perdedor del sorteo de los alumnos, como una variable instrumental para estimar el tratamiento en los tratados (TOT), la recepción real de la beca. Por último, pudieron efectuar un análisis de costo-beneficio para entender mejor el impacto del programa de vales tanto en los gastos de los hogares como del gobierno. Los investigadores llegaron a la conclusión de que los costos sociales totales del programa eran pequeños y se veían compensados por los retornos previstos para los participantes y sus familias, lo que sugiere que los programas orientados a la demanda, como PACES, pueden ser una solución costo-efectiva para aumentar los logros educativos.

Fuente: Angrist et al. (2002).

su control. Concretamente, la parte de la decisión que puede controlarse es la asignación a los grupos de tratamiento y comparación. En la medida en que la asignación a los grupos de tratamiento y de comparación predice la inscripción final en el programa, la asignación aleatoria se puede usar como un "instrumento" para predecir la inscripción final. Tener esta VI permite recuperar las estimaciones del LATE de las estimaciones del efecto de ITT para el tipo de unidades *Inscritos si se lo ofrecen*.

Una variable VI debe satisfacer dos condiciones básicas:

1. No debería estar correlacionada con las características de los grupos de tratamiento y de comparación. Esto se consigue asignando el tratamiento

aleatoriamente a las unidades en la muestra de evaluación. Esto se conoce como *exogeneidad*. Es importante que la VI no influya directamente en el resultado de interés. Los impactos deben ser causados únicamente a través del programa que nos interesa evaluar.

2. Debe influir en las tasas de participación de los grupos de tratamiento y comparación de manera diferente. Normalmente se piensa en aumentar la participación en el grupo de tratamiento. Esto se puede verificar constatando que la participación es más alta en el grupo de tratamiento que en el de comparación. Esta condición se conoce como *relevancia*.

Interpretación de la estimación del efecto promedio del tratamiento local

La diferencia entre la estimación de un ATE y la estimación de un LATE es especialmente importante cuando se trata de interpretar los resultados de una evaluación. Piénsese sistemáticamente en cómo interpretar una estimación LATE. En primer lugar, debe reconocerse que los individuos que cumplen en un programa (el tipo *Inscritos si se lo ofrecen*) son diferentes de los individuos que no cumplen (los tipos *Nuncas* y *Siempres*). Concretamente, en el grupo de tratamiento, los no cumplidores/no participantes (*Nuncas*) pueden ser aquellos que esperan ganar poco con la intervención. En el grupo de comparación, los no cumplidores/participantes (*Siempres*) probablemente constituyan el grupo de individuos que esperan el mayor beneficio de participar. En el ejemplo de la formación docente, los maestros asignados a la capacitación pero que deciden no participar (el tipo *Nuncas*) pueden ser aquellos que creen que no necesitan formación, maestros con un mayor costo de oportunidad del tiempo (por ejemplo, porque tienen un segundo empleo o porque tienen que cuidar de sus hijos), o maestros regidos por una supervisión laxa, que pueden dejar de asistir sin tener problemas. Por otro lado, los docentes asignados al grupo de comparación pero que se inscriben de todas maneras (el tipo *Siempres*) pueden ser aquellos que creen que necesitan formación, maestros que no tienen hijos que cuidar o maestros con un director estricto que insiste en que todos tienen que recibir capacitación.

En segundo lugar, se sabe que la estimación LATE proporciona el impacto para un subgrupo particular de la población: tiene en cuenta solo al subgrupo que no se ve afectado por ningún tipo de incumplimiento. En otras palabras, tiene en cuenta solo el tipo *Inscritos si se lo ofrecen*. Dado que el tipo *Inscritos si se lo ofrecen* es diferente de los *Nuncas* y de los *Siempres*, el impacto que se halla a través de la estimación LATE no se aplica a los tipos *Nuncas o Siempres*. Por ejemplo, si el ministerio de

Educación decidiera implementar una segunda ronda de capacitación y pudiera obligar a los maestros *Nuncas* que no recibieron formación en la primera ronda a recibirla en esta ocasión, no se sabe si esos maestros tendrían efectos menores, iguales o mayores en comparación con los participantes de la primera ronda. De la misma manera, si los docentes más auto motivados siempre encuentran una manera de seguir la capacitándose a pesar de ser asignados de forma aleatoria al grupo de comparación, el LATE para los cumplidores de los grupos tanto de tratamiento como de comparación no proporciona información acerca del impacto del programa para los maestros sumamente motivados (los *Siempres*). La estimación del LATE se aplica únicamente a un subconjunto específico de la población, a saber, aquellos tipos que no están afectados por la falta de cumplimiento –es decir, solo el tipo cumplidor– y no debería extrapolarse a otros subconjuntos de la población.

Promoción aleatoria como variable instrumental

En la sección anterior, se expuso cómo estimar el impacto sobre la base de la asignación aleatoria del tratamiento, aun cuando el cumplimiento con los grupos de tratamiento y comparación originalmente asignados sea imperfecto. A continuación, se propone un enfoque muy similar que se puede aplicar a la evaluación de programas que tienen elegibilidad universal o inscripción abierta, o en los que el administrador del programa no puede controlar quién participa y quién no.

Este enfoque, denominado *promoción aleatoria*, proporciona un estímulo más para que un conjunto aleatorio de unidades se inscriba en el programa. Esta promoción aleatoria sirve como VI. Sirve como una fuente externa de variación que afecta la probabilidad de recibir tratamiento, pero no está relacionada de ninguna forma con las características de los participantes.

Los programas de participación voluntaria suelen permitir que los individuos que se interesan en el programa decidan por sí mismos si quieren inscribirse y participar. Piénsese una vez más en el programa de formación laboral tratado anteriormente, aunque esta vez la asignación aleatoria no es posible y cualquier individuo que desee inscribirse en el programa puede hacerlo. De manera muy parecida a la del ejemplo anterior, se prevé encontrar diferentes tipos de personas: cumplidores, un grupo de *Siempres* y un grupo de *Nuncas*.

- *Siempres*. Los individuos que siempre se inscribirán en el programa.
- *Nuncas*. Los individuos que jamás se inscribirán.

- *Cumplidores o Inscritos si se promueve.* En este contexto cualquier individuo que quiera inscribirse en el programa puede hacerlo. Sin embargo, algunos individuos pueden estar interesados en inscribirse, pero por diversos motivos no tienen suficiente información o el incentivo correcto para hacerlo. En este caso, los cumplidores son aquellos que *se inscriben si se promueve*. Se trata de un grupo de individuos que se anotan en el programa solo si se les ofrece un incentivo adicional, un estímulo o motivación que los impulse a participar. Sin este estímulo adicional, los *Inscritos si se promueve* sencillamente quedarían fuera del programa.

Para volver al ejemplo de la formación para el empleo, si la agencia que organiza la capacitación está bien financiada y tiene suficiente capacidad, puede que despliegue una política de "puertas abiertas" y trate a todas las personas desempleadas que quieran participar. Sin embargo, es poco probable que todas las personas desempleadas quieran participar o incluso que sepan que el programa existe. Puede que algunas personas desempleadas tengan reparos para inscribirse porque saben muy poco acerca del contenido de la formación y les cuesta obtener información adicional. Supóngase que la agencia de formación para el empleo contrata a un trabajador de extensión comunitaria para que se pasee por la ciudad a fin de alentar a un grupo de personas desempleadas seleccionado de forma aleatoria para que se inscriban en el programa de formación laboral. Con la lista de personas desempleadas elegidas de manera aleatoria, llama a sus puertas, describe el programa de formación y les ofrece ayuda para inscribirse en ese mismo momento. La visita es una forma de promoción o estímulo para participar en el programa. Desde luego, no se puede obligar a nadie a participar. Además, las personas desempleadas que el trabajador de extensión comunitaria no visita también pueden inscribirse, aunque tendrán que ir personalmente a la agencia para hacerlo. Por lo tanto, ahora hay dos grupos de personas desempleadas: aquellas que fueron asignadas de modo aleatorio a una visita del trabajador comunitario y aquellas que aleatoriamente no fueron visitadas. Si el esfuerzo de extensión es efectivo, la tasa de inscripción entre las personas desempleadas que fueron visitadas debería ser superior a la tasa entre las personas desempleadas que no fueron visitadas.

Piénsese ahora en cómo se puede evaluar este programa de formación laboral. No se puede simplemente comparar a las personas desempleadas que se inscriben con aquellas que no se inscriben. Esto se debe a que los desempleados que se inscriben probablemente sean muy diferentes de aquellos que no lo hacen, tanto en sus características observables como no observables. Puede que tengan un nivel educativo mayor o menor (esto puede observarse con facilidad) y probablemente estén más motivados y deseosos de encontrar un empleo (esto es difícil de observar y medir).

Sin embargo, hay una variación adicional que se puede explotar para encontrar un grupo de comparación válido. Piénsese si se puede comparar el grupo de personas que fueron asignadas aleatoriamente para recibir una visita del trabajador de extensión con el grupo que no fue visitado. Dado que los grupos con promoción y sin promoción fueron determinados de forma aleatoria, ambos contienen composiciones idénticas de personas muy motivadas (*Siempres*) que se inscribirán independientemente de que el trabajador de extensión llame a su puerta o no. Ambos grupos también contienen personas no motivadas (*Nuncas*) que no se inscribirán en el programa, a pesar de los esfuerzos del trabajador de extensión. Por último, si el trabajador de extensión es efectivo motivando a las personas a inscribirse, algunos (*Inscritos si se promueve*) se anotarán en el programa si el trabajador de extensión los visita, pero no lo harán si no reciben dicha visita.

Dado que el trabajador de extensión visitó a un grupo de individuos asignados de manera aleatoria, puede derivarse una estimación LATE, como se señalaba anteriormente. La única diferencia es que en lugar de asignar el programa de modo aleatorio, se lo está promoviendo aleatoriamente. Siempre que los *Inscritos si se promueve* (que se inscriben cuando se hace contacto con ellos pero no se anotan si no hay contacto) sean lo suficientemente numerosos, entre el grupo *con* la promoción y el grupo *sin* la promoción habrá variaciones que permitirán identificar el impacto de la formación en los *Inscritos si se promueve*. En lugar de cumplir la asignación del tratamiento, los *Inscritos si se promueve* ahora cumplen con la promoción.

Para que esta estrategia funcione, la actividad de promoción tiene que ser efectiva y aumentar la inscripción considerablemente en el grupo de *Inscritos si se promueve*. Al mismo tiempo, las actividades de promoción en sí mismas no deberían influir en los resultados finales de interés (como los ingresos), dado que al final lo que interesa sobre todo es estimar el impacto del programa de formación y no el impacto de la estrategia de promoción en los resultados finales. Por ejemplo, si el trabajador de extensión ofreció grandes cantidades de dinero a los desempleados para conseguir que se inscribieran, sería difícil saber si algún cambio posterior en los ingresos fue causado por la formación o por la actividad de promoción.

La promoción aleatoria es una estrategia creativa que genera el equivalente de un grupo de comparación para los fines de la evaluación de impacto. Se puede usar cuando un programa tiene inscripción abierta y es posible organizar una campaña de promoción destinada a una muestra aleatoria de la población de interés. La promoción aleatoria es otro ejemplo de VI que permite evaluar el impacto de manera no sesgada. Sin embargo, una vez más, como sucede con la asignación aleatoria con cumplimiento imperfecto, las evaluaciones de impacto que dependen de la promoción aleatoria proporcionan una

estimación LATE: una estimación local del efecto en un subgrupo específico de la población, el grupo de *Inscritos si se promueve*. Como sucedió antes, esta estimación LATE no puede extrapolarse directamente al conjunto de la población, dado que los grupos de *Siempres* y *Nuncas* probablemente sean bastante diferentes del grupo de *Inscritos si se promueve*.

¿Ha dicho "promoción"?

La promoción aleatoria pretende aumentar la aceptación de un programa voluntario en una submuestra de la población seleccionada aleatoriamente. La promoción puede adoptar diversas formas. Por ejemplo, puede que se decida iniciar una campaña de información para llegar a aquellas personas que no se han inscrito porque no lo sabían o porque no entienden cabalmente el contenido del programa. También, se pueden ofrecer incentivos para inscribirse, como pequeños obsequios o premios, o facilitando el transporte.

Como se señaló de manera más general en el caso de las VI, para que el método de promoción aleatoria genere una estimación válida del impacto del programa debe cumplirse una serie de condiciones:

1. Los grupos que son objeto y no objeto de la promoción deben ser similares. Es decir, las características promedio de los dos grupos deben ser estadísticamente equivalentes. Esto se consigue asignando de forma aleatoria las actividades de extensión o promoción entre las unidades de la muestra de evaluación.

2. La propia promoción no debería influir directamente en los resultados de interés. Este es un requisito crítico, de modo que se pueda saber que los cambios en los resultados de interés son provocados por el programa mismo y no por la promoción.

3. La campaña de promoción debe alterar considerablemente las tasas de inscripción en el grupo objeto de la promoción en relación con el grupo que no ha sido objeto de la misma. Normalmente, se piensa en aumentar la inscripción mediante la promoción. Esto se puede verificar constatando que las tasas de inscripción sean más altas en el grupo que es objeto de la promoción que en el grupo que no lo es.

Concepto clave

La promoción aleatoria es un método de variables instrumentales que permite estimar el impacto de manera no sesgada. Asigna aleatoriamente una promoción o incentivo para participar en el programa. Es una estrategia útil para evaluar programas que están abiertos a todos los que sean elegibles.

El proceso de promoción aleatoria

El proceso de promoción aleatoria se presenta en el gráfico 5.3. Al igual que con los métodos anteriores, se comienza con la población de unidades elegibles para el programa. A diferencia de la asignación aleatoria, ya no se puede elegir aleatoriamente quién recibirá el programa y quién no lo

Gráfico 5.3 Proceso de promoción aleatoria

recibirá, porque el programa es totalmente voluntario. Sin embargo, en la población de unidades elegibles, habrá tres tipos de unidades:

- *Siempres*. Aquellos que siempre quieren inscribirse en el programa.
- *Inscritos si se promueve*. Aquellos que se inscriben en el programa solo si son objeto de la promoción.
- *Nuncas*. Aquellos que jamás se inscriben en el programa, independientemente de que sean objeto de la promoción o no.

Una vez más, nótese que ser un *Siempre*, un *Inscrito si se promueve* o un *Nunca* es una característica intrínseca de las unidades que no se puede medir fácilmente con un equipo de evaluación del programa porque está relacionado con factores como la motivación, la inteligencia y la información.

Una vez que se define la población elegible, el paso siguiente consiste en seleccionar de manera aleatoria una muestra de la población que formará parte de la evaluación. Estas son las unidades sobre las que se recopilan datos. En algunos casos, por ejemplo, cuando se dispone de datos sobre todas las unidades elegibles, se puede incluir al conjunto de la población en la muestra de evaluación.

Una vez que se ha definido la muestra de evaluación, la promoción aleatoria asigna aleatoriamente la muestra mencionada en el grupo objeto de la promoción y un grupo que no es objeto de ella. Dado que tanto los miembros del grupo con promoción como los del grupo sin promoción se escogen de forma aleatoria, ambos grupos compartirán las

características de la muestra de evaluación general, que serán equivalentes a las características de la población de las unidades elegibles. Por lo tanto, el grupo que es objeto de la promoción y el grupo que no lo es tendrán características similares.

Después de acabar la campaña de promoción, pueden observarse las tasas de inscripción en ambos grupos. En el grupo sin promoción, se inscribirán solo los *Siempres*. Aunque se sabe qué unidades son *Siempres* en el grupo sin promoción, en este grupo no se podrá distinguir entre los *Nuncas* y los *Inscritos si se promueve*. En cambio, en el grupo con promoción se inscribirán tanto los *Inscritos si se promueve* como los *Siempres*, mientras que los *Nuncas* no se anotarán. Por ello, en el grupo con promoción se podrá identificar al grupo de *Nuncas*, pero no se podrá distinguir entre los *Inscritos si se promueve* y los *Siempres*.

Estimación de impacto bajo la promoción aleatoria

Imagínese que en un grupo de 10 individuos la campaña de promoción aumenta la inscripción de un 30% en el grupo sin promoción (3 *Siempres*) a un 80% en el grupo con promoción (3 *Siempres* y 5 *Inscritos si se promueve*). Supóngase que el resultado promedio de todos los individuos del grupo sin promoción (10 individuos) es 70, y que el resultado promedio de los individuos en el grupo con promoción (10 individuos) es 110 (gráfico 5.4). ¿Cuál sería el impacto del programa?

Gráfico 5.4 Estimación del efecto local promedio del tratamiento bajo la promoción aleatoria

	Grupo con promoción	Grupo sin promoción	Impacto
	Porcentaje de inscritos = 80% Media *Y* para grupo con promoción = 110	Porcentaje de inscritos = 30% Media *Y* para grupo sin promoción = 70	Δ porcentaje de inscritos = 50% Δ *Y* = 40 LATE = 40%/50% = 80
Nunca			—
Inscrito si es objeto de la promoción			
Siempre			—

Nota: Las figuras que aparecen con el fondo sombreado corresponden a los que se inscriben. Δ = impacto causal; *Y* = resultado.

En primer lugar, calcúlese la simple diferencia de los resultados entre los grupos con promoción y sin promoción, que es de 40 (110 - 70). Se sabe que ningún elemento de esta diferencia de 40 proviene de los *Nuncas* porque estos no se inscriben en ningún grupo. También se sabe que ningún elemento de la diferencia de 40 se debe a los *Siempres* porque estos se inscriben en ambos grupos. Por lo tanto, toda la diferencia de 40 tendría que deberse al grupo *Inscritos si se promueve*.

El segundo paso consiste en obtener la estimación LATE del programa de los *Inscritos si se promueve*. Se sabe que la diferencia de 40 entre los grupos con promoción y sin promoción puede atribuirse a los *Inscritos si se promueve*, que constituyen solo el 50% de la población. Para evaluar el efecto promedio del programa en un cumplidor, se divide 40 por el porcentaje de *Inscritos si se promueve* en la población. Aunque no se puede identificar directamente a los *Inscritos si se promueve*, se puede deducir cuál debe ser su porcentaje de la población, es decir, la diferencia en las tasas de inscripción de los grupos con promoción y sin promoción (50%, o 0,5). Por lo tanto, la estimación del efecto local promedio del tratamiento del programa del grupo *Inscritos si se promueve* es 40/0,5 = 80.

Dado que la promoción se asigna de forma aleatoria, los grupos con promoción y sin promoción tienen iguales características. Por lo tanto, las diferencias que se observan en los resultados promedio entre los dos grupos tienen que deberse al hecho de que en el grupo con promoción los *Inscritos si se promueve* se inscriben, mientras que en el grupo sin promoción no lo hacen. Una vez más, los impactos estimados de los *Inscritos si se promueve* no deberían extrapolarse directamente a otros grupos, puesto que es probable que sean bastante diferentes de los grupos que se inscriben *Nunca* y *Siempre*. El recuadro 5.3 presenta un ejemplo de promoción aleatoria para un proyecto en Bolivia.

Recuadro 5.3: Promoción de inversiones en infraestructura educativa en Bolivia

En 1991 Bolivia institucionalizó y amplió un exitoso Fondo de Inversión Social (FIS) que ofrecía financiamiento a comunidades rurales para infraestructura de educación, salud y agua. El Banco Mundial, que contribuía al financiamiento del fondo, incorporó una evaluación de impacto al diseño del programa.

Como parte de la evaluación de impacto del componente educativo, se seleccionaron aleatoriamente comunidades de la región de Chaco para la promoción activa del FIS.

Continúa en la página siguiente.

Recuadro 5.3: Promoción de inversiones en infraestructura educativa en Bolivia *(continúa)*

Estas recibieron visitas adicionales de incentivos para participar. El programa estaba abierto a todas las comunidades elegibles en la región y estaba orientado a la demanda, ya que las comunidades debían solicitar fondos para un proyecto específico. La participación fue mayor entre las comunidades con promoción.

Newman et al. (2002) usaron la promoción aleatoria como variable instrumental. Observaron que las inversiones en educación lograron mejorar la calidad de ciertos aspectos de la infraestructura escolar, como la electricidad, las instalaciones de saneamiento, el número de libros de texto por estudiante y la proporción de estudiantes por profesor. Sin embargo, detectaron un escaso impacto en los resultados educativos, con la excepción de un descenso de alrededor del 2,5% en la tasa de abandono escolar. Como consecuencia de estas observaciones, el Ministerio de Educación y el FIS dedican ahora más atención y recursos al *software* de la educación, y solo financian mejoras de la infraestructura física cuando forman parte de una intervención integral.

Fuente: Newman et al. (2002).

Evaluación de impacto del HISP: promoción aleatoria

A continuación, se procurará utilizar el método de promoción aleatoria para evaluar el impacto del HISP. Supóngase que el ministerio de Salud toma la decisión ejecutiva de que el subsidio de seguro de salud debería estar disponible inmediatamente para cualquier hogar que quiera inscribirse. Nótese que se trata de un escenario diferente del caso de asignación aleatoria que se ha estudiado hasta ahora. Sin embargo, usted sabe que en términos realistas esta ampliación a nivel nacional será progresiva lo largo del tiempo, de modo que llega a un acuerdo para intentar acelerar la inscripción en un subconjunto aleatorio de pueblos mediante una campaña de promoción. En una submuestra aleatoria de los pueblos, usted emprende un esfuerzo intensivo de promoción que incluye la comunicación y el marketing social con el fin de crear conciencia de la existencia del HISP. Las actividades de promoción están diseñadas cuidadosamente para evitar contenidos que puedan incentivar de forma involuntaria cambios en otros comportamientos relacionados con la salud, dado que esto invalidaría la promoción como VI. En cambio, la promoción se concentra exclusivamente en aumentar la inscripción en el HISP. Después de dos años de promoción y de implementación del programa, se observa que el 49,2% de los hogares de los pueblos que fueron

asignados aleatoriamente a la promoción se ha inscrito en el programa, mientras que solo lo ha hecho un 8,4% de los hogares de los pueblos sin promoción (cuadro 5.1).

Dado que los pueblos con promoción y sin promoción fueron asignados aleatoriamente, se sabe que las características promedio de los dos grupos deberían ser las mismas en ausencia de la promoción. Dicho supuesto puede verificarse comparando los gastos básicos en salud (así como cualquier otra característica) de las dos poblaciones. Después de dos años de implementación del programa, se observa que el gasto promedio en salud en los pueblos con promoción es de US$14,97 versus US$18,85 en las zonas sin promoción (una diferencia inferior a US$3,87). Sin embargo, dado que la única diferencia entre los pueblos con promoción y sin promoción es que la inscripción en el programa ha sido más alta en los pueblos con promoción (gracias a la campaña de promoción), esta diferencia de US$3,87 en gastos de salud tiene que deberse al 40,78% adicional de hogares que se inscribieron en los pueblos con promoción debido precisamente a la promoción. Por lo tanto, hay que ajustar las diferencias en gastos sanitarios para encontrar el impacto del programa en los *Inscritos si se promueve*. Para esto, se divide la estimación de la ITT –es decir, la simple diferencia entre los grupos con promoción y sin promoción– por el porcentaje de *Inscritos si se promueve*: −3,87/0,4078 = US$9,49. Su colega, un especialista en econometría que sugiere utilizar la promoción aleatoria como variable instrumental, estima el impacto del programa mediante un procedimiento de mínimos cuadrados en dos etapas (véase el manual técnico en línea en

Cuadro 5.1 Evaluación del HISP según la promoción aleatoria (comparación de medias)

	Pueblos con promoción	Pueblos sin promoción	Diferencia	*t*-estadístico
Línea de base: gasto en salud de los hogares	17,19	17,24	−0,05	−0,47
Encuesta de seguimiento: gasto en salud de los hogares	14,97	18,85	−3,87	−16,43
Participación en el HISP	49,20%	8,42%	40,78%	49,85

** Significativo al nivel del 1%.

Cuadro 5.2 Evaluación del HISP según la promoción aleatoria (análisis de regresión)

	Regresión lineal	Regresión lineal multivariante
Impacto estimado sobre el gasto en salud de los hogares	−9,50** (0,52)	−9,74** (0,46)

Nota: Los errores estándares se encuentran entre paréntesis.

** Significativo al nivel del 1%.

www.worldbank.org/ieinpractice para más detalles sobre el enfoque econométrico para estimar los impactos con VI). Su colega encuentra los resultados que aparecen en el cuadro 5.2. Este impacto estimado es válido para aquellos hogares que se inscribieron en el programa debido a la promoción, pero que de otra manera no se habrían inscrito: en otras palabras, los *Inscritos si se promueve*.

Pregunta HISP 4

A. ¿Cuáles son las condiciones clave requeridas para aceptar los resultados de la evaluación de promoción aleatoria del HISP?

B. Sobre la base de estos resultados, ¿se debería ampliar el HISP a nivel nacional?

Limitaciones del método de promoción aleatoria

La promoción aleatoria es una estrategia útil para evaluar el impacto de programas voluntarios y programas con elegibilidad universal, sobre todo porque no requiere la exclusión de ninguna unidad elegible. Sin embargo, el enfoque tiene algunas limitaciones en comparación con la asignación aleatoria del tratamiento.

En primer lugar, la estrategia de promoción debe ser efectiva. Si la campaña de promoción no aumenta la inscripción, no aparecerá ninguna diferencia entre los grupos con promoción y sin promoción, y no habrá nada que comparar. Por lo tanto, es crucial diseñar cuidadosamente la campaña de promoción y realizar una prueba piloto extensiva de la misma para asegurarse de que será efectiva. El aspecto positivo es que el diseño de dicha campaña puede ayudar a los administradores del programa enseñándoles cómo aumentar la inscripción después de que haya concluido el período de evaluación.

En segundo lugar, el método de promoción aleatoria permite estimar el impacto del programa solo para un subconjunto de la población de unidades

elegibles (un LATE). Concretamente, el impacto promedio local del programa se estima a partir del grupo de individuos que se inscriben únicamente cuando se les incentiva a hacerlo. Sin embargo, puede que los individuos de este grupo tengan características muy diferentes de aquellos que siempre se inscriben o nunca se inscriben. Por lo tanto, el efecto promedio del tratamiento para el conjunto de la población puede ser distinto del efecto del tratamiento promedio estimado para los individuos que participan solo cuando se les incentiva. Una evaluación con promoción aleatoria no estimará los impactos en el grupo de individuos que se inscriben en el programa sin ser incentivados. En algunos casos, este grupo (los *Siempres*) puede ser precisamente el grupo que el programa está diseñado para beneficiar. En este contexto, el diseño de promoción aleatoria arrojará luz sobre los impactos esperados en nuevas poblaciones que se inscribirían debido a la promoción adicional, pero no en cuanto a la población que ya se ha inscrito por su propia iniciativa.

Lista de verificación: promoción aleatoria como variable instrumental

La promoción aleatoria genera estimaciones válidas del contrafactual si la campaña de promoción aumenta de forma considerable la aceptación del programa sin influir directamente en los resultados de interés.

- ✓ Las características de línea de base, ¿están equilibradas entre las unidades que recibieron la campaña de promoción y aquellas que no la recibieron? Compárense las características de línea de base de los dos grupos.
- ✓ La campaña de promoción, ¿ha influido de forma considerable en la aceptación del programa? Tendría que influir. Compárense las tasas de aceptación del programa en las submuestras con promoción y sin promoción.
- ✓ La campaña de promoción, ¿influye directamente en los resultados? No tendría que influir. Esto no puede comprobarse directamente, de modo que tiene que depender de la teoría, del sentido común y del conocimiento adecuado del entorno de la evaluación de impacto como guía.

Recursos adicionales

- Para material de apoyo del libro e hipervínculos a recursos adicionales, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para otros recursos sobre VI, véase el portal de evaluación del Banco Interamericano de Desarrollo (BID) (http//:www.iadb.org/portalevaluacion).

Notas

1. En la ciencia médica, los pacientes del grupo de comparación suelen recibir un placebo, es decir, una píldora edulcorada que no tendrá efecto en el resultado previsto. Esto se hace con el fin de controlar mejor el *efecto placebo*, es decir, los cambios potenciales en la conducta y los resultados que podrían darse sencillamente por el acto de recibir un tratamiento, aunque el tratamiento mismo sea inefectivo.
2. Estos dos pasos corresponden a la técnica econométrica de mínimos cuadrados en dos etapas, que produce una estimación del efecto local promedio del tratamiento.

Referencias bibliográficas

Angrist, J., E. Bettinger, E. Bloom, E. King y M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (5): 1535–58.

Kearney, M. S. y P. B. Levine. 2015. "Early Childhood Education by MOOC: Lessons from Sesame Street." Documento de trabajo NBER 21229, National Bureau of Economic Research, Cambridge, MA.

Newman, J., M. Pradhan, L. B. Rawlings, G. Ridder, R. Coa y J. L. Evia. 2002. "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund." *World Bank Economic Review* 16 (2): 241–74.

CAPÍTULO 6

Diseño de regresión discontinua

Evaluación de programas que utilizan un índice de elegibilidad

Los programas sociales a menudo utilizan un índice para decidir quién tiene derecho a inscribirse en un programa y quién no. Por ejemplo, los programas de lucha contra la pobreza suelen focalizarse en los hogares pobres, identificados mediante una puntuación o un índice de la pobreza. El índice de pobreza se puede basar en una fórmula que mide un conjunto de activos básicos del hogar como factor aproximado (o estimativo) de sus medios (como el ingreso, el consumo o el poder adquisitivo).[1] Se clasifica a los hogares con baja puntuación como pobres, y a los hogares con puntuaciones más altas se les considera relativamente acomodados. Los programas de lucha contra la pobreza suelen establecer un umbral o una puntuación límite, por debajo del cual se determina la condición de pobreza y la elegibilidad para el programa. El sistema de selección de beneficiarios del gasto social en Colombia es un ejemplo de este tipo de esquema (véase el recuadro 6.1). Las puntuaciones en las pruebas educativas constituyen otro ejemplo (recuadro 6.3). Se puede conceder la admisión a la universidad a los individuos que obtienen los mejores resultados en las pruebas estandarizadas, calificados desde el más bajo al más alto. Si el número de becas es limitado, solo serán admitidos los alumnos con puntuaciones por encima de un cierto umbral (como, por ejemplo, el 10% superior de los alumnos). En ambos ejemplos hay un índice de

Recuadro 6.1: Uso del diseño de regresión discontinua para evaluar el impacto de la reducción de las tarifas escolares en los índices de matriculación en Colombia

Barrera-Osorio, Linden y Urquiola (2007) utilizaron un diseño de regresión discontinua (DRD) para evaluar el impacto de un programa para reducir las tarifas escolares en Colombia (Gratuidad) sobre los índices de matriculación en los colegios de la ciudad de Bogotá. El programa definió su población objetivo sobre la base del SISBEN, un índice continuo de pobreza cuyo valor está determinado por características de los hogares, como la ubicación, los materiales de construcción de la vivienda, los servicios de los que dispone, la demografía, la salud, la educación, el ingreso y las ocupaciones de los miembros de la familia. El gobierno estableció dos puntuaciones límite en el índice SISBEN. Así, los niños de los hogares con puntuaciones inferiores a la primera puntuación límite eran elegibles para recibir educación gratuita entre los grados 1 y 11, los niños de los hogares cuyas puntuaciones se hallaban entre la primera y la segunda puntuación eran elegibles para un subsidio del 50% en las tarifas para los grados 10 y 11, y los niños de los hogares con puntuaciones superiores a la segunda puntuación no eran elegibles para recibir educación gratuita ni subsidios.

Los autores utilizaron un DRD por cuatro motivos. En primer lugar, las características del hogar, como el ingreso o el nivel educativo del jefe de familia, son continuos en la puntuación SISBEN en la línea de base; en otras palabras, no hay "saltos" en las características en la puntuación SISBEN. En segundo lugar, los hogares en ambos lados de las puntuaciones límite tienen características similares, y generan grupos de comparación creíbles. En tercer lugar, se disponía de una muestra grande de hogares. Por último, el gobierno mantuvo en secreto la fórmula utilizada para calcular el índice SISBEN, de modo que no se pudieran manipular las puntuaciones.

Al usar el método DRD, los investigadores observaron que el programa tuvo un impacto positivo significativo en los índices de matriculación escolar. Concretamente, la matriculación fue 3 puntos porcentuales más alta en los alumnos de primaria de hogares con puntuaciones inferiores a la primera puntuación límite y 6 puntos porcentuales más alta en los alumnos de secundaria de los hogares ubicados entre la primera y la segunda puntuación límite. Este estudio aporta evidencia sobre los beneficios de reducir los costos directos de la escolarización, sobre todo entre los alumnos en situación de riesgo. Sin embargo, sus autores también recomiendan una investigación más detallada sobre las elasticidades con respecto al precio para fundamentar mejor el diseño de los programas de subsidios como este.

Fuente: Barrera-Osorio, Linden y Urquiola (2007).

elegibilidad continuo, así como también un umbral o una puntuación límite que determina quién es elegible y quién no lo es.

El *diseño de regresión discontinua* (DRD) es un método de evaluación de impacto que se puede utilizar en programas que tienen un índice de elegibilidad continuo con un umbral (puntuación límite) de elegibilidad definido con claridad para determinar quién es elegible y quién no lo es. A fin de aplicar un DRD, deben cumplirse las siguientes condiciones:

1. El índice debe clasificar a las personas o unidades de una manera continua o "fluida". Índices como el de pobreza, las puntuaciones de las pruebas estandarizadas o la edad tienen numerosos valores que se pueden ordenar de menor a mayor y, por lo tanto, se pueden considerar continuos. En cambio, las variables con categorías discretas que solo tienen unos pocos valores posibles o no se pueden ordenar, no se consideran continuas. Ejemplos de esta última clase son la condición laboral (empleado o desempleado), el nivel más alto de estudios alcanzado (primario, secundario, universitario o posgrado), la propiedad de un automóvil (sí o no) o el país de nacimiento.

2. El índice debe tener una puntuación límite claramente definida, es decir, un punto por debajo o por encima del cual se clasifica a la población como elegible para el programa. Por ejemplo, los hogares con un índice de pobreza igual o menor a 50 sobre 100 se podrían clasificar como pobres, los individuos mayores de 67 años se podrían clasificar como elegibles para una jubilación, y los alumnos con una puntuación superior a 90 sobre 100 podrían considerarse elegibles para una beca. Las puntuaciones límite en estos ejemplos son 50, 67 y 90 respectivamente.

3. La puntuación límite debe ser única para el programa de interés, es decir, aparte del programa que se evalúa, no debería haber otros programas que utilicen la misma puntuación límite. Por ejemplo, si un índice de pobreza por debajo de 50 clasifica a un hogar para recibir una transferencia de efectivo, un seguro de salud y transporte público gratis, no se podría utilizar el método DRD para estimar por sí solo el impacto del programa de transferencias de efectivo.

4. La puntuación de un individuo o una unidad particular no puede ser manipulada por los encuestadores, los beneficiarios potenciales, los administradores del programa o los políticos.

El DRD estima el impacto en torno a la puntuación límite de elegibilidad como la diferencia entre el resultado promedio de unidades del lado tratado de la puntuación límite de elegibilidad y el resultado promedio de unidades en el lado no tratado (comparación) de la puntuación límite.

Concepto clave

El diseño de regresión discontinua (DRD) es un método de evaluación de impacto adecuado para programas que utilizan un índice continuo para clasificar a los participantes potenciales y que tienen una puntuación límite en el índice que determina si los participantes potenciales tienen derecho o no a beneficiarse del programa.

Piénsese en un programa agrícola cuyo fin es aumentar la producción total de arroz subsidiando la compra de fertilizantes de los agricultores. El programa está destinado a fincas pequeñas y medianas con menos de 50 hectáreas. Antes del comienzo del programa, se puede esperar que las fincas más pequeñas tengan una producción menor que las grandes, como se muestra en el gráfico 6.1, que ilustra el tamaño de la finca y la producción de arroz. En este caso, la puntuación de elegibilidad es el número de hectáreas de la finca, y la puntuación límite es de 50 hectáreas. Las reglas del programa establecen que las fincas por debajo de 50 hectáreas son elegibles para recibir subsidios para fertilizantes, y las fincas de 50 o más hectáreas no lo son. Entonces, se puede prever que participará del programa una cantidad de fincas de 48, 49 o incluso 49,9 hectáreas. Y habrá otro grupo con 50, 50,1 y 50,2 hectáreas que no participará del programa, porque esas fincas superan la puntuación límite. Es probable que el grupo de fincas con 49,9 hectáreas sea muy similar al grupo de aquellas que tienen 50,1 hectáreas en todos los aspectos, salvo que un grupo recibió el subsidio para fertilizantes y el otro no. A medida que nos alejamos de la puntuación límite de elegibilidad, hay más diferencias entre las fincas elegibles. Sin embargo, la extensión de las fincas es una buena medida de sus diferencias, y permite controlar por una buena parte de esas diferencias.

Gráfico 6.1 Producción de arroz, fincas pequeñas vs. fincas grandes (línea de base)

Una vez que el programa se pone en marcha y subvenciona el costo del fertilizante de las fincas pequeñas y medianas, la evaluación de impacto podría utilizar un DRD para evaluar su impacto (gráfico 6.2). El DRD calcula el impacto como la diferencia en los resultados, por ejemplo, de la producción de arroz, entre las unidades a ambos lados del límite de elegibilidad, que en este caso es un tamaño de finca de 50 hectáreas. Las fincas que eran demasiado grandes para inscribirse en el programa constituyen el grupo de comparación y generan una estimación del resultado contrafactual de esas fincas del grupo de tratamiento que eran justo lo suficientemente pequeñas para inscribirse. Dado que estos dos grupos eran muy similares en la línea de base y están expuestos al mismo conjunto de factores externos a lo largo del tiempo (como el clima, los shocks de precios y las políticas agrícolas locales y nacionales), el único motivo plausible de la diferencia en los resultados tiene que ser el propio programa.

Dado que el grupo de comparación está compuesto por fincas que superan la puntuación de elegibilidad, el impacto dado por un DRD es válido solo "a nivel local", es decir, en la cercanía del límite de elegibilidad. De esta manera, se obtiene una estimación de un efecto local promedio del tratamiento (LATE) (véase el capítulo 5). El impacto del programa de subvenciones de fertilizantes es válido para las fincas más grandes

Gráfico 6.2 Producción de arroz, fincas pequeñas vs. fincas grandes (seguimiento)

dentro de aquellas de tamaño medio, es decir, aquellas cuya extensión se halla justo por debajo de las 50 hectáreas. La evaluación de impacto no será necesariamente capaz de identificar directamente el impacto del programa en las fincas más pequeñas –por ejemplo, las de 10 o 20 hectáreas de superficie–, donde los efectos de una subvención de los fertilizantes podrían diferir en aspectos importantes de las fincas de tamaño medio de 48 o 49 hectáreas. Una ventaja del método DRD es que una vez que se aplican las reglas de elegibilidad del programa, no es preciso dejar a ninguna unidad elegible sin tratamiento para los fines de la evaluación de impacto. La otra cara de la moneda es que los impactos de las observaciones lejos de la puntuación límite no se conocerán. El recuadro 6.2 presenta un ejemplo del uso del DRD para evaluar un programa de redes de protección social en Jamaica.

Recuadro 6.2: Redes de protección social basadas en un índice de pobreza en Jamaica

El método de diseño de regresión discontinua (DRD) se utilizó para evaluar el impacto de una iniciativa de redes de protección social en Jamaica. En 2001 el gobierno de este país lanzó el programa Advancement through Health and Education (PATH) (Salud y Educación para el Progreso) destinado a aumentar las inversiones en capital humano y mejorar la focalización de los beneficios de las prestaciones sociales para los pobres. El programa ofrecía subsidios de salud y educación a los niños de los hogares pobres elegibles, con la condición de que asistieran a la escuela y realizaran chequeos médicos de manera regular. El beneficio mensual promedio para cada niño fue de alrededor de US$6,50, además de una exención estatal de ciertas tasas en salud y educación.

Después de determinar la elegibilidad para el programa con una fórmula de puntuación, Levy y Ohls (2010) pudieron comparar los hogares justo por debajo del umbral de elegibilidad con los hogares justo por encima (con una diferencia de entre 2 y 15 puntos con respecto a la puntuación límite). Los investigadores justificaron el uso del método de DRD con datos de línea de base que mostraban que los hogares de tratamiento y comparación tenían niveles similares de pobreza, medidos por las puntuaciones de tipo "*proxy mean*", y niveles similares de motivación, dado que todos los hogares de la muestra habían postulado al programa. Los investigadores también utilizaron la puntuación de elegibilidad del programa en el análisis de regresión para controlar por cualquier diferencia entre ambos grupos.

Levy y Ohls (2010) llegaron a la conclusión de que el programa PATH aumentaba la asistencia escolar de los niños de entre 6 y

Continúa en la página siguiente.

Recuadro 6.2: Redes de protección social basadas en un índice de pobreza en Jamaica *(continúa)*

17 años en una media de 0,5 días al mes, lo cual es significativo, dado que la tasa de asistencia ya era bastante alta (85%). Además, las visitas a los centros de salud de niños de 0 a 6 años aumentaron en alrededor de un 38%. Aunque los investigadores no pudieron encontrar ningún impacto de más largo plazo en los logros escolares ni en la condición de salud, llegaron a la conclusión de que la magnitud de los impactos que hallaron era en general consistente con los programas de transferencias condicionadas implementados en otros países. Un aspecto final interesante de esta evaluación es que recopiló tanto datos cuantitativos como cualitativos, utilizando sistemas de información, entrevistas, grupos focales y encuestas de hogares.

Fuente: Levy y Ohls (2010).

El diseño de regresión discontinua difuso

Aun cuando se haya verificado que no existe evidencia de manipulación en el índice de elegibilidad, puede que todavía persista un problema si las unidades no respetan su asignación al grupo de tratamiento o de comparación. En otras palabras, algunas unidades que tienen derecho al programa sobre la base de su índice de elegibilidad pueden decidir no participar, mientras que otras unidades que no tenían derecho al programa sobre la base de su índice de elegibilidad pueden encontrar un modo de participar de todas maneras. Cuando todas las unidades cumplen con la asignación que les corresponde sobre la base de su índice de elegibilidad, se dice que el DRD es "nítido", y si hay incumplimiento en alguno de los lados de la puntuación límite, se dice que el DRD es "difuso" (gráfico 6.3). Si el DRD es difuso, se puede utilizar el enfoque de variable instrumental para corregir por la falta de cumplimiento (véase el capítulo 5). Recuérdese que en el caso de la asignación aleatoria con incumplimiento, se utiliza la asignación aleatoria como la variable instrumental que ayudó a corregir por la falta de cumplimiento. En el caso del DRD, se puede usar la asignación original basada en el índice de elegibilidad como variable instrumental. Sin embargo, hacerlo tiene un inconveniente, a saber, que la estimación de impacto con el DRD instrumental será más localizada en el sentido de que ya no es válida para todas las observaciones cercanas a la puntuación límite sino que representa el impacto para el subgrupo de la población situada cerca de la puntuación límite y que participa en el programa solo debido a los criterios de elegibilidad.

Gráfico 6.3 Cumplimiento de la asignación

Verificación de la validez del diseño de regresión discontinua

Para que un DRD produzca una estimación LATE no sesgada de la puntuación límite, es importante que el índice de elegibilidad no sea manipulado en la cercanía de la puntuación límite de modo que un individuo pueda cambiar su condición de tratamiento o control.[2] La manipulación de los criterios de elegibilidad puede adoptar numerosas formas. Por ejemplo, los encuestadores que recopilan los datos que se utilizan para calcular la puntuación de elegibilidad podrían cambiar una o dos respuestas de los encuestados; o puede que los encuestados mientan deliberadamente a los encuestadores si creen que con eso tendrán acceso al programa. Además, la manipulación de las puntuaciones puede agravarse a lo largo del tiempo, a medida que los encuestadores, los encuestados y los políticos comienzan a aprender las "reglas del juego". En el ejemplo de la subvención de los fertilizantes, la manipulación en torno al límite se produciría si los agricultores pudieran alterar los títulos de propiedad o si dieran informes falsos sobre el tamaño de sus fincas. O un agricultor con 50,3 hectáreas de tierra podría encontrar una manera de vender media hectárea para ser elegible para el programa, en el caso de que los beneficios previstos de la subvención a los fertilizantes merecieran la pena hacerlo.

Una de las señales que delata la manipulación se ilustra en el gráfico 6.4. El panel (a) muestra la distribución de los hogares según su índice de línea de base cuando no hay manipulación. La densidad de los hogares en torno al límite (50) es continua (o fluida). El panel (b) presenta una situación diferente: un número mayor de hogares parecen estar agrupados justo por

Gráfico 6.4 Manipulación del índice de elegibilidad

Recuadro 6.3: El efecto en el desempeño escolar de la agrupación de alumnos según sus puntuaciones en las pruebas educativas en Kenia

Para comprobar si la asignación de alumnos a clases sobre la base del desempeño mejora los resultados educativos, Duflo, Dupas y Kemer (2011) llevaron a cabo un experimento con 121 escuelas primarias en la región occidental de Kenia. En la mitad de las escuelas, los alumnos de primer grado fueron distribuidos de forma aleatoria en dos diferentes secciones de la clase. En la otra mitad de los colegios, los alumnos fueron asignados a una sección de alto o de bajo desempeño según sus puntuaciones en las pruebas iniciales, usando la puntuación de las pruebas educativas como punto límite.

El diseño de regresión discontinua (DRD) les permitió comprobar a los autores si la composición de los alumnos de una clase tenía un efecto directo en las puntuaciones de las pruebas. Los investigadores compararon las puntuaciones finales de las pruebas de los alumnos justo en torno al límite para ver si aquellos asignados a la sección de alto desempeño tenían mejores resultados que aquellos asignados a la sección de bajo desempeño.

En promedio, las puntuaciones de las pruebas finales de los colegios que agruparon estudiantes en clases con niveles similares de desempeño fueron 0,14 desviaciones estándar más altas que en el caso de los colegios que no usaron este método y en cambio utilizaron la asignación aleatoria para crear grupos equivalentes de estudiantes. Estos resultados no fueron solo producto de los alumnos en las secciones de alto desempeño, dado que los estudiantes de la sección de bajo desempeño también mostraron mejoras en las puntuaciones de las pruebas. En el caso de los alumnos justo en torno a la puntuación límite, los investigadores encontraron que no había una diferencia significativa en las puntuaciones finales de las pruebas. Estas conclusiones rechazan la hipótesis de que los alumnos se benefician directamente al tener compañeros de clase con un desempeño superior.

Fuente: Duflo, Dupas y Kemer (2011).

debajo del límite, mientras que hay relativamente pocos hogares justo por encima del límite. Dado que no hay un motivo a priori para creer que debería haber un gran cambio en el número de hogares justo en torno al límite, la ocurrencia de ese cambio en la distribución en torno al límite es una prueba de que de alguna manera los hogares pueden estar manipulando sus puntuaciones para tener acceso al programa. Una segunda prueba de manipulación grafica el índice de elegibilidad en relación con la variable de resultado en la línea de base y verifica que no haya discontinuidad, o un "salto", justo en torno a la línea del límite.

Evaluación de impacto del HISP: diseño de regresión discontinua

Piénsese en cómo se puede aplicar el método de diseño de regresión discontinua (DRD) al Programa de Subsidios de Seguros de Salud (HISP). Después de llevar a cabo investigaciones sobre el diseño del HISP, se descubre que además de seleccionar aleatoriamente los pueblos de tratamiento, las autoridades localizaron el programa en los hogares de bajos ingresos utilizando la línea nacional de pobreza. La línea de la pobreza se basa en un índice de pobreza que asigna a cada hogar en el país una puntuación entre 20 y 100 en función de sus activos, las condiciones de la vivienda y la estructura sociodemográfica. La línea de pobreza ha sido fijada oficialmente en 58. Esto significa que todos los hogares con una puntuación de 58 o menos se clasifican como pobres, y que todos los hogares con una puntuación de más de 58 se consideran no pobres. Incluso en los pueblos de tratamiento, solo los hogares pobres son elegibles para inscribirse en el HISP. La base de datos con la que se cuenta contiene información tanto de los hogares pobres como de los no pobres en las comunidades de tratamiento.

Antes de llevar a cabo las estimaciones del diseño de regresión discontinua, se decide verificar si hay evidencia de manipulación del índice de elegibilidad. Como primera medida, se verifica si la densidad del índice de elegibilidad suscita alguna preocupación a propósito de la manipulación del índice. Luego se grafica el porcentaje de hogares en contraste con el índice de pobreza de la línea de base (gráfico 6.5).[3] El gráfico no señala ninguna "concentración" de los hogares justo por debajo del límite de 58.

A continuación, se verifica si los hogares respetaron su asignación a los grupos de tratamiento y comparación sobre la base de su puntuación de elegibilidad. Se gráfica la participación en el programa en contraste con el índice de pobreza de línea de base (gráfico 6.6) y se observa que

Gráfico 6.5 HISP: densidad de los hogares, según el índice de pobreza de línea de base

Gráfico 6.6 Participación en el HISP, según el índice de pobreza de línea de base

dos años después del comienzo del plan piloto, solo los hogares con una puntuación de 58 o menos (es decir, a la izquierda de la línea de la pobreza) han podido inscribirse en el HISP. Además, todos los hogares elegibles se inscribieron en el HISP. En otras palabras, se observa un cumplimiento total, por lo cual se obtiene un DRD "nítido".

A continuación, se procede a aplicar el método de DRD para calcular el impacto del programa. Utilizando datos del seguimiento, se vuelve a graficar la relación entre las puntuaciones en el índice de pobreza y los gastos previstos en salud, y se observa la relación que se ilustra en el gráfico 6.7. En la relación entre el índice de pobreza y los gastos previstos en salud, se encuentra una clara ruptura, o *discontinuidad* de la línea de pobreza (58).

La discontinuidad refleja una disminución de los gastos en salud en aquellos hogares elegibles para beneficiarse del programa. Dado que los hogares en ambos lados de la puntuación límite de 58 son muy similares, la explicación plausible de la diferencia en el nivel de gastos en salud es que un grupo de los hogares era elegible para inscribirse en el programa y el otro no. Esta diferencia se estima a través de una regresión con los datos recogidos en el cuadro 6.1.

Gráfico 6.7 Índice de pobreza y gastos en salud: el HISP dos años después

Cuadro 6.1 Evaluación del HISP: diseño de regresión discontinua con análisis de regresión

	Regresión lineal multivariante
Impacto estimado en el gasto en salud de los hogares	–9,03**
	(0,43)

Nota: Los errores estándar están entre paréntesis. ** Significativo al nivel del 1%.

Pregunta HISP 5

A. El resultado que se refleja en el cuadro 6.1, ¿es válido para todos los hogares elegibles?

B. En comparación con el impacto estimado con la asignación aleatoria, ¿qué dice este resultado con respecto a los hogares con un índice de pobreza justo por debajo de 58?

C. De acuerdo con las estimaciones de impacto del DRD, ¿se debería ampliar el HISP a nivel nacional?

Limitaciones e interpretaciones del método de diseño de regresión discontinua

El diseño de regresión discontinua proporciona estimaciones del LATE en torno al límite de elegibilidad en el punto en que las unidades de tratamiento y comparación son más similares. Cuanto más se acerque uno a la puntuación límite, más similares serán las unidades a cada lado del umbral. De hecho, si uno se aproxima mucho a la puntuación límite, las unidades a ambos lados del umbral serán tan similares que su comparación será tan buena como si los grupos de tratamiento y de comparación se hubieran elegido mediante la asignación aleatoria del tratamiento.

Dado que el método de DRD estima el impacto del programa en torno a la puntuación límite, o *localmente*, la estimación no se puede necesariamente generalizar a unidades cuyas puntuaciones se alejan más del umbral, es decir, donde los individuos elegibles y no elegibles quizá no sean tan similares. El hecho de que el método de DRD no pueda proporcionar una estimación de un efecto de tratamiento promedio para todos los participantes del programa puede verse a la vez como una fortaleza y como una limitación, lo cual depende de la pregunta de la evaluación de interés. Si la

evaluación pretende responder la pregunta ¿el programa debería existir o no?, el efecto de tratamiento promedio para toda la población elegible puede ser el parámetro más relevante, y es evidente que el DRD no será del todo perfecto. Sin embargo, si la cuestión de interés para las políticas es ¿el programa debería suspenderse o ampliarse? –es decir, para los beneficiarios (potenciales) justo en las inmediaciones del límite–, el DRD produce precisamente la estimación local de interés para sustentar esta importante decisión de política.

Como ya se señaló, pueden surgir otras complicaciones cuando el cumplimiento en cualquiera de los dos lados del límite es imperfecto. Este DRD difuso se produce cuando las unidades que no son elegibles debido a su puntuación en el índice consiguen tener acceso al programa, o cuando las unidades elegibles según la puntuación del índice deciden no participar en el programa. En este caso, se puede utilizar una metodología de variable instrumental similar a la definida en el capítulo 5, a saber, la localización de las unidades por encima o por debajo de la puntuación límite se usará como variable instrumental para la participación observada en el programa. Como sucedía en los ejemplos del capítulo 5, esto tiene un inconveniente: solo se puede estimar el impacto de aquellas unidades que son "sensibles" al criterio de elegibilidad, esto es si se trata del tipo *Inscrito si es elegible*, pero no si se trata del tipo *Siempres* o *Nuncas*.

El hecho de que el método de DRD estime el impacto solo en las inmediaciones de las puntuaciones límite también genera dificultades en términos de la potencia estadística del análisis. En ocasiones, solo se emplea en el análisis un conjunto limitado de observaciones que se sitúan cerca de la puntuación límite, con lo cual el número de observaciones en el análisis de DRD se reduce, en comparación con los métodos que analizan todas las unidades en los grupos de tratamiento y comparación. Para obtener una potencia estadística suficiente al aplicar el DRD, habrá que utilizar un *ancho de banda* en torno a la puntuación límite que incluya un número suficiente de observaciones. En la práctica, se debería intentar utilizar un ancho de banda lo más amplio posible, a la vez que se conserva el equilibrio en las características observadas de la población por encima y por debajo de la puntuación límite. Luego, se puede aplicar la estimación varias veces usando diferentes anchos de banda para verificar si las estimaciones son sensibles al ancho de banda utilizado.

Es necesario formular otra advertencia al utilizar el método de DRD, a saber, la especificación puede ser sensible a la forma funcional que se emplea para modelar la relación entre la puntuación de elegibilidad y el resultado de interés. En los ejemplos presentados en este capítulo, se da por sentado que la relación entre el índice de elegibilidad y el resultado es lineal. En realidad, la relación podría ser más compleja, e incluir relaciones no lineales e

interacciones entre variables. Si uno no se da cuenta de estas relaciones complejas en la estimación, se las puede confundir con una discontinuidad, lo que llevaría a una interpretación incorrecta de la estimación de impacto con DRD. En la práctica, se puede estimar el impacto del programa utilizando diversas formas funcionales (lineales, cuadráticas, cúbicas, cuárticas, y otras similares) para evaluar si, de hecho, las estimaciones de impacto son sensibles a la forma funcional.

Por último, como se señala más arriba, hay unas cuantas condiciones importantes para la regla de elegibilidad y el umbral. En primer lugar, deben ser únicos del programa de interés. Por ejemplo, puede utilizarse un índice de pobreza que establezca un *ranking* de hogares o individuos para focalizar una diversidad de programas sociales para los pobres. En este caso, no será posible aislar el impacto de un solo programa de lucha específica contra la pobreza de todos los demás programas que utilizan los mismos criterios de focalización. En segundo lugar, la regla de elegibilidad y el umbral deberían ser resistentes a la manipulación de los encuestadores, los beneficiarios potenciales, los administradores de los programas o los políticos. La manipulación del índice de elegibilidad crea una discontinuidad en el índice que socava la condición básica para que el método funcione, a saber, que el índice de elegibilidad debería ser continuo en torno al umbral.

Incluso con estas limitaciones, el DRD es un poderoso método de evaluación de impacto para generar estimaciones no sesgadas del impacto de un programa en la cercanía del límite de elegibilidad. El DRD aprovecha las reglas de asignación del programa, a partir de índices de elegibilidad continuos, que ya son habituales en numerosos programas sociales. Cuando se aplican las reglas de focalización basadas en el índice, no es necesario excluir un grupo de hogares o individuos elegibles como beneficiarios del tratamiento a los fines de la evaluación, porque se puede utilizar el diseño de regresión discontinua como alternativa.

Lista de verificación: diseño de regresión discontinua

El DRD requiere que el índice de elegibilidad sea continuo en torno a la puntuación límite, y que las unidades sean similares en las cercanías por encima o por debajo de la puntuación límite.

- ✓ ¿Es continuo el índice en torno la puntuación límite en el momento de la línea de base?
- ✓ ¿Hay alguna evidencia de falta de cumplimiento de la regla que determine la elegibilidad para el tratamiento? Compruébese que todas las unidades

elegibles y ninguna unidad no elegible han recibido el tratamiento. Si se encuentra falta de cumplimiento, habrá que combinar el DRD con un enfoque de variable instrumental para corregir esta "discontinuidad difusa".[4]

- ✓ ¿Hay alguna evidencia de que las puntuaciones del índice puedan haber sido manipuladas con el fin de influir en quien tenía derecho a beneficiarse del programa? Compruébese si la distribución de la puntuación del índice es fluida en el punto límite. Si se halla evidencia de una "concentración" de puntuaciones ya sea por encima o por debajo del punto límite, puede que esto sea una señal de manipulación.
- ✓ ¿El umbral corresponde a un único programa que se está evaluando o está siendo usado por otros programas también?

Otros recursos

- Para material de apoyo de este libro y para hipervínculos de otros recursos, se recomienda consultar el sitio web de Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para información acerca de la evaluación de un programa de transferencias de efectivo utilizando un DRD, véase la entrada en el blog de impacto del desarrollo del Banco Mundial http://blogs.worldbank.org/impactevaluations/.
- Para una revisión de los temas prácticos en la implementación del DRD, véase G. Imbens y T. Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.

Notas

1. En ocasiones, esto se denomina prueba de medias *proxy*.
2. El índice de elegibilidad continuo a veces se denomina variable forzada.
3. Nota técnica: la densidad se estimó utilizando el método de estimación univariante del núcleo de Epanechnikov.
4. En este caso, se utilizaría la localización a la izquierda o la derecha del punto límite como variable instrumental para la aceptación del programa en la práctica en la primera etapa de una estimación de mínimos cuadrados en dos etapas.

Referencias bibliográficas

Barrera-Osorio, F., L. Linden y M. Urquiola. 2007. "The Effects of User Fee Reductions on Enrollment: Evidence from a Quasi-Experiment." Washington, D.C.: Columbia University y Banco Mundial.

Duflo, E., P. Dupas y M. Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101: 1739–74.

Imbens, G. y T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.

Levy, D. y J. Ohls. 2010. "Evaluation of Jamaica's PATH Conditional Cash Transfer Programme." *Journal of Development Effectiveness* 2 (4): 421–41.

CAPÍTULO 7

Diferencias en diferencias

Evaluación de un programa cuando la regla de asignación es menos clara

Los tres métodos de evaluación de impacto analizados hasta ahora, asignación aleatoria, variables instrumentales (VI) y diseño de regresión discontinua (DRD), estiman el contrafactual mediante reglas explícitas de asignación del programa que el equipo de evaluación conoce y entiende. Se ha visto por qué estos métodos ofrecen estimaciones creíbles del contrafactual haciendo relativamente pocas suposiciones e imponiendo pocas condiciones. Los dos próximos tipos de métodos, *diferencias en diferencias* (DD) y *pareamiento*, ofrecen al equipo de evaluación herramientas adicionales que pueden aplicarse cuando las reglas de asignación del programa son menos claras o cuando no es factible ninguno de los tres métodos antes descritos. En casos como este, se suele utilizar tanto el método de diferencias en diferencias como el de pareamiento. Sin embargo, ambos también requieren supuestos más fuertes que la asignación aleatoria, o los métodos de VI o DRD. Se entiende que si no se conoce la regla de asignación del programa, hay una incógnita más en la evaluación, acerca de la cual se deben formular supuestos. Dado que los supuestos no son necesariamente verdaderos, puede que el método de diferencias en diferencias o el de pareamiento no siempre proporcionen estimaciones fiables de los impactos de los programas.

El método de diferencias en diferencias

Concepto clave

El método de *diferencias en diferencias* compara los cambios en los resultados a lo largo del tiempo entre unidades inscritas en un programa (el grupo de tratamiento) y unidades que no lo están (el grupo de comparación). Esto permite corregir cualquier diferencia entre los grupos de tratamiento y comparación que sea constante a lo largo del tiempo.

El método de *diferencias en diferencias* contrasta las *diferencias* en los resultados a lo largo del tiempo entre una población inscrita en un programa (el grupo de tratamiento) y una población no inscrita (el grupo de comparación). Piénsese, por ejemplo, en un programa de reparación de carreteras que se lleva a cabo a nivel de distrito pero que no se puede asignar aleatoriamente entre distritos ni se asigna sobre la base de un índice con un umbral claramente definido, lo cual permitiría un diseño de regresión discontinua. Las juntas de los distritos pueden decidir inscribirse o no inscribirse en el programa. Uno de los objetivos del programa consiste en mejorar el acceso de la población a los mercados laborales, y uno de los indicadores de resultados es la tasa de empleo. Como se analizó en el capítulo 3, el solo hecho de observar el cambio antes y después en las tasas de empleo en los distritos que se inscriben en el programa no capturará el impacto causal del programa porque es probable que muchos otros factores influyan en el empleo a lo largo del tiempo. A la vez, comparar los distritos que se inscribieron y los que no se inscribieron en el programa de reparación de carreteras será problemático si existen motivos no observables por los que algunos distritos se inscribieron en el programa y otros no lo hicieron (el problema de sesgo de selección en el escenario de inscritos versus no inscritos).

Sin embargo, ¿qué pasaría si se combinan los dos métodos y se comparan los cambios antes-después en los resultados de un grupo que se inscribió en el programa con los cambios antes-después de un grupo que no se inscribió en el programa? La diferencia en los resultados antes-después para el grupo inscrito –*la primera diferencia*– controla por factores que son constantes a lo largo del tiempo en ese grupo, puesto que se está comparando el propio grupo consigo mismo. Sin embargo, todavía quedan los factores externos que varían con el tiempo (*factores variables en el tiempo*) en este grupo. Una manera de capturar esos factores que varían en el tiempo es medir el cambio antes-después en los resultados de un grupo que no se inscribió en el programa pero que estuvo expuesto al mismo conjunto de condiciones ambientales (*la segunda diferencia*). Si se "limpia" la primera diferencia de otros factores variables en el tiempo que influyen en el resultado de interés sustrayendo la segunda diferencia, se habrá eliminado una fuente de sesgo que resultaba preocupante en las comparaciones sencillas antes-después. El enfoque de diferencias en diferencias hace lo que su nombre sugiere: combina las dos estimaciones falsas del contrafactual (comparaciones antes-después y comparaciones entre quienes se inscriben y quienes deciden no hacerlo) para producir una mejor estimación del contrafactual. En el ejemplo del programa de reparación de carreteras, el método DD podría comparar los cambios en el empleo antes y después de que se ponga en marcha el

programa en los individuos que viven en distritos que lo introdujeron con los cambios en el empleo en los distritos donde no se implementó el programa.

Es importante señalar que el contrafactual que se estima en este caso es el *cambio* en los resultados del grupo de tratamiento. La estimación de este contrafactual es el cambio en los resultados del grupo de comparación. Los grupos de tratamiento y comparación no tienen necesariamente que tener las mismas condiciones antes de la intervención. Sin embargo, para que el método de diferencias en diferencias sea válido, el grupo de comparación debe mostrar con precisión el cambio en los resultados que habría experimentado el grupo de tratamiento en ausencia de tratamiento. Para aplicar diferencias en diferencias, hay que medir los resultados en el grupo que se beneficia del programa (el grupo de tratamiento) con los resultados del grupo que no se beneficia (el grupo de comparación), tanto antes como después del programa. En el recuadro 7.1, se presenta un ejemplo de utilización del método DD para entender el impacto de los incentivos electorales en la implementación de un programa de transferencias condicionadas aplicado en Brasil y en las tasas de deserción escolar.

Recuadro 7.1: Utilización del método DD para entender el impacto de los incentivos electorales en las tasas de abandono escolar en Brasil

En un estudio empírico sobre los incentivos electorales locales, De Janvry, Finan y Sadoulet (2011) analizan los impactos de un programa de transferencias condicionadas en Brasil. El programa Bolsa Escola entregaba a las madres de los hogares pobres una mensualidad con la condición de que sus hijos asistieran a la escuela. Se trataba de un programa federal similar al de Oportunidades de México (véanse los recuadros 1.1 y 4.2), pero a nivel municipal. Las municipalidades eran las encargadas de identificar a los beneficiarios e implementar el programa.

Utilizando el método de diferencias en diferencias, los autores estimaron el impacto del programa en las tasas de abandono escolar, y encontraron una variación notable en el desempeño del programa en las diferentes municipalidades. Para explorar esta variación, los investigadores compararon la mejora en las tasas de abandono escolar en los municipios cuyos alcaldes ejercían su primer mandato con los municipios donde los alcaldes ya estaban en su segundo mandato. Su hipótesis era que, dado que en Brasil existe un límite de dos mandatos para los cargos locales, a los alcaldes que se hallaban en su primer mandato les preocupaba su reelección y, por lo tanto, actuaban de manera diferente que los alcaldes que ejercían ya en su segundo mandato, que no tenían esas preocupaciones.

En general, el programa tuvo éxito y redujo las tasas de abandono escolar en un promedio

Continúa en la página siguiente.

Recuadro 7.1 Utilización del método DD para entender el impacto de los incentivos electorales en las tasas de abandono escolar en Brasil *(continúa)*

del 8% entre los beneficiarios. Los investigadores observaron que el impacto del programa era un 36% mayor en los municipios cuyos alcaldes ejercían su primer mandato. Su conclusión fue que las preocupaciones acerca de la reelección incentivaron a los políticos locales a aumentar sus esfuerzos en la implementación del programa Bolsa Escola.

Fuente: De Janvry, Finan y Sadoulet (2011).

Gráfico 7.1 El método de diferencias en diferencias

Nota: Todas las diferencias entre los puntos deberían leerse como diferencias verticales de los resultados en el eje vertical.

El gráfico 7.1 ilustra el método de diferencias en diferencias en el ejemplo de la reparación de carreteras. El año 0 es el año de línea de base. En el año 1 se inscribe en el programa un grupo de distritos de tratamiento, mientras que no lo hace un grupo de distritos de comparación. El nivel de los resultados (la tasa de empleo) en el grupo de tratamiento va de *A*, antes del comienzo del programa, a *B*, después del comienzo del programa, mientras que el resultado para el grupo de comparación va de *C*, antes del comienzo del programa, a *D*, después del comienzo del programa.

Recuérdense los dos falsos contrafactuales: la diferencia de los resultados antes y después de la intervención en el grupo de tratamiento (*B*-*A*) y la diferencia de los resultados después de la intervención entre los grupos de tratamiento y de comparación (*B*–*D*). Con las diferencias en diferencias, la

estimación del contrafactual se obtiene calculando el cambio en los resultados del grupo de comparación (*D*–*C*) y luego sustrayéndolo del cambio en los resultados del grupo de tratamiento (*B*–*A*). Utilizar el cambio en los resultados del grupo de comparación como la estimación del contrafactual para el cambio en los resultados del grupo de tratamiento es similar a suponer que si el grupo que se inscribió no hubiera participado en el programa, sus resultados habrían evolucionado a lo largo del tiempo siguiendo la misma tendencia que el grupo que no se inscribió, es decir, la evolución en el resultado del grupo inscrito habría ido de *A* a *E*, como se muestra en el gráfico 7.1.

En resumen, el impacto del programa se computa simplemente como la diferencia entre dos diferencias:

$$\text{Impacto de la DD} = (B - A) - (D - C) = (0{,}74 - 0{,}60) - (0{,}81 - 0{,}78) = 0{,}11.$$

Las relaciones que se muestran en el gráfico 7.1 también pueden presentarse en un cuadro sencillo. El cuadro 7.1 describe los componentes de las estimaciones de diferencias en diferencias. La primera línea contiene los resultados del grupo de tratamiento antes de la intervención (*A*) y después de la intervención (*B*). La comparación antes-después del grupo de tratamiento es la primera diferencia (*B*–*A*). La segunda línea contiene los resultados del grupo de comparación antes de la intervención (*C*) y después de la intervención (*D*), de modo que la segunda diferencia es (*D*–*C*).

El método de *diferencias en diferencias* computa la estimación del impacto de la siguiente manera:

1. Se calcula la diferencia del resultado (*Y*) entre las situaciones antes y después para el grupo de tratamiento (*B* – *A*).

Cuadro 7.1 Cálculo del método de diferencias en diferencias

	Después	Antes	Diferencia
Tratamiento/inscritos	*B*	*A*	*B – A*
Comparación/no inscritos	*D*	*C*	*D – C*
Diferencia	*B – D*	*A – C*	*DD = (B – A) – (D – C)*

	Después	Antes	Diferencia
Tratamiento/inscritos	0,74	0,60	*0,14*
Comparación/no inscritos	0,81	0,78	*0,03*
Diferencia	*–0,07*	*–0,18*	*DD = 0,14 – 0,03 = 0,11*

2. Se calcula la diferencia del resultado *(Y)* entre las situaciones antes y después para el grupo de comparación *(D – C)*.

3. A continuación, se calcula la diferencia entre la diferencia en los resultados del grupo de tratamiento *(B – A)* y la diferencia del grupo de comparación *(D – C)*, o *DD* = *(B – A) – (D – C)*. Estas diferencias en diferencias constituyen la estimación del impacto.

También se consideran las diferencias en diferencias en la dirección contraria: se calcula primero la diferencia en el resultado entre el grupo de tratamiento y el de comparación en la situación después; luego se calcula la diferencia en el resultado entre el grupo de tratamiento y de comparación en la situación antes, y finalmente se sustrae este último del primero.

$$\text{Impacto DD} = (B - D) - (A - C) = (0{,}74 - 0{,}81) - (0{,}60 - 0{,}78) = 0{,}11.$$

¿Qué utilidad tiene el método de diferencias en diferencias?

Para entender la utilidad de este método, debe tomarse nuestro segundo contrafactual falso, analizado en el capítulo 3, que comparaba las unidades inscritas con las no inscritas en un programa. Recuérdese que la principal preocupación en este caso era que las dos series de unidades pudieran tener características diferentes y que pueden ser dichas características –y no el programa– las que explican la diferencia en los resultados entre los dos grupos. Las diferencias *no observadas* en las características eran especialmente preocupantes: por definición, es imposible incluir las características no observables en el análisis.

El método de diferencias en diferencias contribuye a resolver este problema en la medida en que se puede razonablemente suponer que muchas características de las unidades o personas son constantes a lo largo del tiempo (o *invariables en el tiempo*). Piénsese, por ejemplo, en características observables, como el año de nacimiento de una persona, la ubicación de una región con respecto al océano, la altura de la ciudad o el nivel de educación de los padres. Es probable que la mayoría de estos tipos de variables, aunque posiblemente relacionadas con los resultados, no cambien en el transcurso de una evaluación. Con el mismo razonamiento, podría llegarse a la conclusión de que muchas características no observables de los individuos también son más o menos constantes a lo largo del tiempo. Piénsese, por ejemplo, en los rasgos de la personalidad o el historial de salud de la familia. Es posible que estas características intrínsecas de las personas no cambien con el tiempo.

En lugar de contrastar los resultados entre los grupos de tratamiento y comparación después de la intervención, los métodos de diferencias en diferencias estudian las *tendencias* entre los grupos de tratamiento y comparación. La tendencia de un individuo es la diferencia en los resultados para ese individuo antes y después del programa. Al sustraer la situación de los resultados *antes* de la situación *después*, se anula el efecto de todas las características que son únicas de ese individuo y que no cambian a lo largo del tiempo. En realidad, se está anulando (o controlando) no solo el efecto de características *observables* invariables en el tiempo, sino también el efecto de características *no observables* invariables en el tiempo, como las ya mencionadas. En el recuadro 7.2 se describe un estudio que utilizó el método de diferencias en diferencias para estimar el impacto de una mayor presencia policial en la incidencia de robos de vehículos en Buenos Aires.

Concepto clave

En lugar de contrastar resultados entre los grupos de tratamiento y comparación después de la intervención, los métodos de diferencias en diferencias comparan las *tendencias* entre ambos grupos.

Recuadro 7.2: Aplicación del método de diferencias en diferencias para estudiar los efectos del despliegue policial en la tasa de delitos en Argentina

DiTella y Schargrodsky (2005) analizaron si un mayor despliegue de las fuerzas policiales reducía los delitos en Argentina. En 1994 un ataque terrorista contra un importante centro judío en Buenos Aires llevó al gobierno argentino a aumentar la protección policial de los edificios relacionados con instituciones judías en el país.

Con el objetivo de entender el impacto de la presencia policial en la incidencia del delito, los autores recopilaron datos sobre el número de robos de vehículos por manzana en tres barrios en Buenos Aires antes y después del ataque terrorista. Luego combinaron esta información con datos geográficos sobre la ubicación de instituciones judías en aquellos barrios. Este estudio presentó un enfoque diferente de las habituales regresiones utilizadas en la lucha contra el crimen. Los trabajos sobre el impacto de la presencia policial a menudo se enfrentan a un problema de endogeneidad, puesto que los gobiernos tienden a aumentar la presencia policial en zonas con tasas de delitos más altas. En cambio, el incremento en el despliegue de la fuerza policial en Argentina no estaba relacionado en absoluto con la incidencia de los robos de vehículos, de modo que el estudio no sufre de este problema de causalidad simultánea. DiTella y Schargrodsky utilizaron el método de diferencias en diferencias para estimar el impacto de la mayor presencia policial en la incidencia de los robos de vehículos.

Los resultados revelaron un efecto disuasorio positivo de la presencia policial en los delitos. Sin embargo, este efecto era localizado. En las manzanas donde había edificios relacionados con instituciones judías que tenían protección policial, los robos de vehículos disminuyeron significativamente en comparación con otras manzanas, a saber, en un 75%. Los investigadores no encontraron impactos en los robos de vehículos a una o dos manzanas de los edificios protegidos.

Fuente: DiTella y Schargrodsky (2005).

El supuesto de "tendencias iguales" en el método de diferencias en diferencias

A pesar de que las diferencias en diferencias permiten tener en cuenta las diferencias entre los grupos de tratamiento y comparación que son constantes a lo largo del tiempo, no ayudan a eliminar las diferencias entre los grupos de tratamiento y de comparación que cambian con el tiempo. En el ejemplo del programa de reparación de carreteras, si las zonas de tratamiento también se benefician de la construcción de un nuevo puerto marítimo al mismo tiempo que se produce la reparación de las carreteras, el efecto de esta última no podrá separarse de la construcción del puerto marítimo utilizando un enfoque de diferencias en diferencias. Para que el método proporcione una estimación válida del contrafactual, se debe suponer que no existen ese tipo de diferencias que varían en el tiempo entre los grupos de tratamiento y comparación.

Otra manera de pensar en esto es que, en ausencia del programa, las diferencias en los resultados entre los grupos de tratamiento y comparación tendrían que evolucionar de forma paralela. Es decir, sin el tratamiento, los resultados tendrían que aumentar o disminuir en la misma medida en ambos grupos; los resultados tienen que mostrar *tendencias iguales en ausencia de tratamiento*.

Desde luego, no hay manera de demostrar que las diferencias entre los grupos de tratamiento y comparación habrían evolucionado de manera paralela en ausencia del programa. El motivo es que no se puede observar qué habría ocurrido con el grupo de tratamiento en ausencia del tratamiento, es decir, no se puede observar el contrafactual.

Por lo tanto, cuando se emplea el método de diferencias en diferencias, se debe *suponer* que, en ausencia del programa, los resultados en el grupo de tratamiento habrían evolucionado de forma paralela con los resultados del grupo de comparación. El gráfico 7.2 ilustra un incumplimiento de este supuesto fundamental. Si las tendencias de los resultados son diferentes para los grupos de tratamiento y de comparación, el efecto estimado de tratamiento obtenido mediante métodos de diferencias en diferencias sería inválido o estaría sesgado. Esto se debe a que la tendencia del grupo de comparación no es una estimación válida de la tendencia contrafactual que habría prevalecido en el grupo de tratamiento en ausencia del programa. Como se muestra en el gráfico 7.2, si en realidad los resultados del grupo de comparación aumentan más lentamente que los resultados del grupo de tratamiento en ausencia del programa, utilizar la tendencia del grupo de comparación como estimación del contrafactual de la tendencia del grupo de tratamiento conduce a una estimación sesgada del impacto del programa. Más concretamente, se estaría sobreestimando el impacto del programa.

Gráfico 7.2 Diferencias en diferencias cuando las tendencias de los resultados son diferentes

Comprobación del supuesto de igualdad de tendencias en el método de diferencias en diferencias

A pesar de que no se puede demostrar, la validez del supuesto fundamental de igualdad de tendencias se puede evaluar. Una primera verificación de validez consiste en contrastar los cambios en los resultados en los grupos de tratamiento y comparación en repetidas ocasiones antes de la implementación del programa. En el programa de reparación de carreteras, esto significa que se mediría el cambio en la tasa de empleo entre los grupos de tratamiento y comparación antes del comienzo del programa, es decir, entre el segundo y el primer año, y entre el primer año y el año cero. Si se ve que los resultados evolucionaban de forma paralela antes del comienzo del programa, es razonable suponer que habrían seguido evolucionando de la misma manera después de la intervención. Para verificar la igualdad de las tendencias antes de la intervención, se requieren al menos dos rondas de observaciones en los grupos de tratamiento y comparación antes del comienzo del programa. Esto significa que la evaluación requeriría tres rondas de observaciones: dos observaciones previas a la intervención para valorar las tendencias anteriores al programa, y al menos una observación posterior a la intervención para evaluar el impacto mediante el método de diferencias en diferencias.

Una segunda manera de comprobar el supuesto de las tendencias iguales sería llevar a cabo lo que se denomina *prueba de placebo*. Para esta prueba, se realiza una segunda estimación de diferencias en diferencias utilizando un grupo de tratamiento "falso", es decir, un grupo que, según lo que el evaluador sabe, no ha sido afectado por el programa. Por ejemplo, se quiere estimar cómo las clases de apoyo para los alumnos del séptimo grado influyen en su probabilidad de asistir a la escuela, y entonces se eligen estudiantes de octavo grado como grupo de comparación. Para comprobar si los estudiantes de séptimo y octavo grado tienen las mismas tendencias en términos de asistencia escolar, podría verificarse que los estudiantes de octavo y de sexto grado tengan las mismas tendencias. El evaluador sabe que los alumnos de sexto grado no se verán afectados por el programa, de modo que si realiza una estimación de diferencias en diferencias utilizando a los alumnos de octavo grado como grupo de comparación y a los de sexto como el grupo de tratamiento falso, *tiene que* obtener un impacto de cero. De lo contrario, el impacto que encuentre se deberá necesariamente a alguna diferencia fundamental en las tendencias entre los estudiantes de sexto y octavo grado. Esto, a su vez, arroja dudas sobre si es válido el supuesto de que los alumnos de séptimo y octavo grado tienen tendencias iguales en ausencia del programa.

Una tercera manera de probar el supuesto de tendencias iguales sería llevar a cabo la prueba de placebo no solo con un grupo de tratamiento falso, sino también con un resultado falso. En el ejemplo de las clases de apoyo, conviene comprobar la validez de utilizar a los alumnos de octavo grado como grupo de comparación estimando el impacto de las clases de apoyo en un resultado que, según lo que se sabe, no se ve afectado por dichas clases, como, por ejemplo, el número de hermanos que los alumnos tienen. Si la estimación de diferencias en diferencias encuentra un "impacto" de las clases de apoyo en el número de hermanos de los alumnos, entonces ya se sabe que el grupo de comparación debe tener alguna falla.

Una cuarta manera de comprobar el supuesto de las tendencias iguales consistiría en aplicar el método de diferencias en diferencias utilizando diferentes grupos de comparación. En el ejemplo de las clases de apoyo, primero se llevaría a cabo la estimación con los alumnos de octavo grado como grupo de comparación, y luego se realizaría una segunda estimación tomando a los alumnos de sexto grado como grupo de comparación. Si los dos grupos son válidos, se observará que el impacto estimado es aproximadamente el mismo en ambos cálculos. En los recuadros 7.3 y 7.4 se presentan dos ejemplos de una evaluación de diferencias en diferencias que utilizan una combinación de estos métodos para probar el supuesto de tendencias iguales.

Recuadro 7.3: Comprobando el supuesto de tendencias iguales: privatización del agua y mortalidad infantil en Argentina

Galiani, Gertler y Schargrodsky (2005) usaron el método de diferencias en diferencias para resolver una importante pregunta de las políticas públicas: ¿la privatización de los servicios de suministro de agua mejora los resultados en materia de salud y contribuye al alivio de la pobreza? Durante la década de 1990, Argentina inició una de las campañas de privatización más grandes de su historia y transfirió las compañías municipales de aguas a empresas privadas reguladas. El proceso de privatización se produjo a lo largo de una década, y el mayor número de privatizaciones tuvo lugar después de 1995, cubriendo alrededor del 30% de los municipios del país y a un 60% de la población.

La evaluación aprovechó el cambio de la propiedad del servicio de aguas a lo largo del tiempo para determinar el impacto de la privatización sobre la tasa de mortalidad de los menores de 5 años. Antes de 1995, el ritmo de reducción de las tasas de mortalidad infantil era el mismo en todo el país; después de 1995, las tasas de mortalidad se redujeron más rápidamente en aquellos municipios donde se había privatizado el suministro de agua.

De acuerdo con los investigadores, en este contexto es muy probable que se cumpla el supuesto necesario para aplicar el método de diferencias en diferencias. Concretamente, los autores demostraron que no se observaban diferencias en las tendencias de mortalidad infantil entre los municipios de comparación y tratamiento antes de que comenzara la campaña de privatizaciones. También demostraron que la decisión de privatizar no guardaba relación alguna con las crisis económicas ni con los niveles históricos de mortalidad infantil. Verificaron la solidez de sus observaciones llevando a cabo una prueba placebo con un "resultado falso": distinguieron entre aquellas causas de mortalidad infantil relacionadas con la calidad del agua, como enfermedades infecciosas y parasitarias, y aquellas no relacionadas con la calidad del agua, como los accidentes y las enfermedades congénitas. Luego probaron el impacto de la privatización de los servicios de suministro de agua por separado para los dos subconjuntos de causas de mortalidad. Así, llegaron a la conclusión de que la privatización de los servicios de suministro de agua estaba correlacionada con la disminución de las muertes por enfermedades infecciosas y parasitarias, pero no estaba correlacionada con la disminución de las muertes por causas como accidentes y enfermedades congénitas.

Al final, la evaluación determinó que la mortalidad infantil se redujo cerca de un 8% en las zonas donde se privatizaron los servicios, y que el efecto fue más notable, de alrededor de un 26%, en las zonas más pobres, donde la ampliación de la red de suministro de agua había sido mayor. Este estudio arrojó luz sobre diversos debates fundamentales de políticas públicas en relación con la privatización de los servicios públicos. Los investigadores llegaron a la conclusión de que en Argentina el sector privado regulado demostraba ser más exitoso que el sector público en la mejora de indicadores de acceso, servicio y, lo que es más importante, mortalidad infantil.

Fuente: Galiani, Gertler y Schargrodsky (2005).

Recuadro 7.4: Poniendo a prueba el supuesto de tendencias iguales: la construcción de escuelas en Indonesia

Duflo (2001) analizó los impactos de mediano y largo plazo de un programa para construir escuelas en Indonesia en los resultados educativos y del mercado laboral. En 1973 Indonesia se embarcó en un programa de construcción de escuelas primarias de gran escala, y se construyeron más de 61.000 establecimientos de este tipo. Para centrarse en los alumnos que no se habían matriculado anteriormente en la escuela, el gobierno asignó el número de colegios que serían construidos en cada distrito en proporción al número de alumnos no matriculados en ese distrito. Duflo deseaba evaluar el impacto del programa en los niveles educativos y los salarios. La exposición al tratamiento se midió por el número de escuelas de la región, y los grupos de tratamiento y comparación fueron identificados a partir de la edad que tenían cuando se lanzó el programa. El grupo de tratamiento se componía de varones nacidos después de 1962, dado que habrían sido lo bastante jóvenes para beneficiarse de las nuevas escuelas primarias cuando estas se construyeron en 1974. El grupo de comparación estaba formado por varones nacidos antes de 1962, que habrían sido demasiado mayores para beneficiarse del programa.

Duflo utilizó el método de diferencias en diferencias para estimar el impacto del programa en los logros educativos promedio de los salarios, y comparó las diferencias en los resultados entre distritos de alta y baja exposición. A fin de demostrar que se trataba de un método de estimación válido, primero tuvo que probar el supuesto de tendencias iguales en diferentes distritos. Para hacerlo, utilizó una prueba de placebo con un grupo de tratamiento falso. Comparó la cohorte de 18 a 24 años en 1974 con la cohorte de 12 a 17 años. Dado que ambas eran demasiado mayores para beneficiarse del nuevo programa, los cambios en sus niveles educativos no deberían ser sistemáticamente diferentes en los distintos distritos. La estimación de esta regresión de diferencias en diferencias era cercana a cero. Este resultado significaba que los niveles educativos antes de que el programa comenzara no aumentaron más rápidamente en las zonas que después se convertirían en distritos de alta exposición que en los distritos de baja exposición. La prueba de placebo también mostró que la estrategia de basarse en la edad en el momento de la construcción de la escuela funcionaría.

La evaluación encontró resultados positivos en los niveles educativos y en los salarios de los alumnos con una alta exposición al programa, es decir, aquellos que tenían menos de 8 años cuando se construyeron las escuelas. Para estos estudiantes, cada nueva escuela construida por cada 1.000 niños se asociaba con una mejora de 0,12 a 0,19 años en los niveles educativos y un aumento del 3% al 5,4% en los salarios. El programa también aumentó en un 12% la probabilidad de que un niño completara la escuela primaria.

Fuente: Duflo (2001).

Evaluación del impacto del HISP: la aplicación de diferencias en diferencias

El método de diferencias en diferencias se puede usar para evaluar el Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés). En este escenario, hay dos rondas de datos sobre dos grupos de hogares: un grupo que se inscribió en el programa y otro que no lo hizo. Si se recuerda el caso de los grupos inscritos y no inscritos, se verá que no se puede simplemente comparar los gastos promedio en salud de los dos grupos debido al sesgo de selección. Puesto que se cuenta con datos de los períodos para cada hogar de la muestra, dichos datos se pueden usar para resolver algunos de estos problemas comparando el cambio en los gastos en salud de ambos grupos, suponiendo que el cambio en el gasto en salud del grupo no inscrito refleje lo que habría ocurrido con los gastos del grupo inscrito en ausencia del programa (véase el cuadro 7.2). Nótese que no importa de qué manera se calcula la doble diferencia.

A continuación, se estima el efecto utilizando análisis de regresión (cuadro 7.3). Recurriendo a una simple regresión lineal para computar la estimación simple de diferencias en diferencias, se observa que el programa redujo los gastos en salud de los hogares en US$ 8,16. Luego se refina el análisis añadiendo otras variables de control. En otras palabras, se emplea una regresión lineal multivariante que tiene en cuenta múltiples otros factores, y se observa la misma reducción en los gastos de los hogares en salud.

Cuadro 7.2 Evaluación del HISP: diferencias en diferencias (comparación de medias)

	Después (seguimiento)	Antes (línea de base)	Diferencia
Inscritos	7,84	14,49	–6,65
No inscritos	22,30	20,79	1,51
Diferencia			DD = –6,65 – 1,51 = –8,16

Nota: El cuadro presenta el gasto medio (en dólares) en salud de los hogares inscritos y no inscritos, antes y después de la introducción del HISP.

Cuadro 7.3 Evaluación del HISP: diferencias en diferencias (análisis de regresión)

	Regresión lineal	Regresión lineal multivariante
Impacto estimado sobre el gasto en salud de los hogares	–8,16** (0,32)	–8,16** (0,32)

Nota: Los errores estándares están entre paréntesis.
** Significativo al nivel del 1%.

Pregunta HISP 6

A. ¿Qué supuestos básicos son necesarios para aceptar este resultado de diferencias en diferencias?

B. De acuerdo con los resultados de las diferencias en diferencias, ¿se debería ampliar el HISP a nivel nacional?

Limitaciones del método de diferencias en diferencias

Aun cuando las tendencias sean iguales antes del comienzo de la intervención, el sesgo en la estimación de diferencias en diferencias puede producirse y pasar inadvertido. Esto se debe a que el método DD atribuye a la intervención cualquier diferencia de las tendencias entre los grupos de tratamiento y de comparación que se producen desde el momento en que la intervención comienza. Si hay otros factores presentes que influyen en la diferencia en las tendencias entre los dos grupos, y la regresión multivariante no rinde cuenta de ellos, la estimación será inválida o sesgada.

Supóngase que se intenta estimar el impacto en la producción de arroz con la subvención de los fertilizantes y que esto se lleva a cabo midiendo la producción de arroz de los agricultores subvencionados (tratamiento) y de los agricultores no subvencionados (comparación) antes y después de la distribución de las subvenciones. Si en el año 1 tiene lugar una sequía que afecta solamente a los agricultores subvencionados, la estimación de diferencias en diferencias producirá una estimación inválida del impacto de subvencionar los fertilizantes. En general, cualquier factor que afecte a uno de los dos grupos de forma desproporcionada, y lo hace al mismo tiempo en que el grupo de tratamiento recibe el tratamiento, sin que esto se tome en cuenta en la regresión, puede potencialmente invalidar o sesgar la estimación del impacto del programa. El método DD supone que no hay factores de este tipo presentes.

Verificación: diferencias en diferencias

Las diferencias en diferencias suponen que las tendencias de los resultados son similares en los grupos de comparación y tratamiento antes de la intervención y que los únicos factores que explican las diferencias en los

resultados entre ambos grupos, aparte del propio programa, son constantes a lo largo del tiempo.

- ✓ Los resultados ¿habrían evolucionado de forma paralela en los grupos de tratamiento y comparación en ausencia del programa? Esto se puede evaluar utilizando diversas pruebas de falsificación, como las siguientes: 1) Los resultados en los grupos de tratamiento y comparación ¿evolucionaban de modo paralelo antes de la intervención? Si hay dos rondas de datos disponibles antes del comienzo del programa, se debe probar si existen diferencias en las tendencias que aparecen entre ambos grupos; 2) ¿Qué sucede con los resultados falsos que no deberían verse afectados por el programa? ¿Evolucionan de forma paralela antes y después del inicio de la intervención en los grupos de tratamiento y comparación?
- ✓ Realizar el análisis de diferencias en diferencias utilizando varios grupos plausibles de comparación. Deberían obtenerse estimaciones similares del impacto del programa.
- ✓ Efectuar el análisis de diferencias en diferencias usando los grupos de tratamiento y comparación elegidos, y un resultado falso que no debería verse afectado por el programa. Debería encontrarse un impacto nulo del programa en ese resultado.
- ✓ Llevar adelante el análisis de diferencias en diferencias utilizando la variable de resultados elegida con dos grupos que, según lo que se sabe, no se vieron afectados por el programa. Debería observarse un impacto cero del programa.

Otros recursos

- Para material de apoyo del libro y para hipervínculos de recursos adicionales, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para más referencias sobre los supuestos no dichos de las diferencias en diferencias, véase la entrada correspondiente en el blog de Impacto del Desarrollo del Banco Mundial (http://blogs.worldbank.org/impactevaluations).

Referencias bibliográficas

De Janvry, A., F. Finan y E. Sadoulet. 2011. "Local Electoral Incentives and Decentralized Program Performance." *The Review of Economics and Statistics* 94 (3): 672–85.

DiTella, R. y E. Schargrodsky. 2005. "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack." *American Economic Review* 94 (1): 115–33.

Duflo, E. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Galiani, S., P. Gertler y E. Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

CAPÍTULO 8

Pareamiento

Construcción de un grupo de comparación artificial

El método que se describe en este capítulo consiste en técnicas estadísticas a las que se denominarán pareamiento (*matching*). Los métodos de pareamiento se pueden aplicar en el contexto de casi todas las reglas de asignación de un programa, siempre que se cuente con un grupo que no haya participado en el mismo. El pareamiento utiliza técnicas estadísticas para construir un grupo de comparación. Para cada unidad posible, el tratamiento intenta encontrar una unidad de no tratamiento (o conjunto de unidades de no tratamiento) que tengan características lo más parecidas posible. Piénsese en un caso en el que se propone evaluar el impacto de un programa de capacitación laboral sobre el ingreso y se cuenta con una base de datos, como los registros de ingreso y las declaraciones tributarias, que contiene tanto a los individuos que se inscribieron en el programa como a los individuos que no lo hicieron. El programa que se intenta evaluar no tiene reglas de asignación claras (como asignación aleatoria o un índice de elegibilidad) que explique por qué ciertos individuos se inscribieron en el programa y otros no lo hicieron. En este contexto, los métodos de pareamiento permitirán identificar el conjunto de individuos no inscritos que más se parece a los individuos tratados, a partir de las características que ya se tienen en la base de datos. Estos individuos no

Concepto clave

El pareamiento utiliza grandes bases de datos y técnicas estadísticas para construir el mejor grupo de comparación posible sobre la base de características observables.

inscritos pareados luego se convierten en el grupo de comparación que se emplea para estimar el contrafactual.

La búsqueda de una buena pareja para cada participante del programa requiere aproximarse todo lo posible a las características que explican la decisión del individuo de inscribirse en el programa. Desafortunadamente, en la práctica esto es más difícil. Si la lista de características observables relevantes es muy grande, o si cada característica adopta muchos valores, puede que sea complicado identificar una pareja para cada una de las unidades del grupo de tratamiento. A medida que aumenta el número de características o dimensiones con las que se quiere parear las unidades que se inscribieron en el programa, puede que uno se encuentre con lo que se denomina la *maldición de la dimensionalidad.* Por ejemplo, si solo se consideran tres características importantes para identificar el grupo de comparación del pareamiento, como la edad, el sexo y si la persona tiene un diploma de estudios secundarios, es probable que se encuentren parejas para todos los participantes que se inscribieron en el programa entre el conjunto de aquellos que no se inscribieron (los no inscritos), pero se corre el riesgo de dejar al margen otras características potencialmente importantes. Sin embargo, si se aumenta la lista de características –por ejemplo, para incluir el número de hijos, el número de años de estudios, el número de meses que el individuo lleva desempleado, el número de años de experiencia, etc.– puede que la base de datos no contenga una buena pareja para la mayoría de los participantes del programa que están inscritos, a menos que abarque un número muy grande de observaciones. El gráfico 8.1 ilustra el pareamiento sobre la base de cuatro características: edad, sexo, meses de desempleo, y diploma de estudios secundarios.

Gráfico 8.1 Pareamiento exacto en cuatro características

Unidades tratadas			
Edad	Género	Meses desempleado	Diploma de secundaria
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Unidades no tratadas			
Edad	Género	Meses desempleado	Diploma de secundaria
24	1	8	1
38	0	1	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

Pareamiento por puntajes de propensión

Por suerte, la maldición de la dimensionalidad puede solucionarse fácilmente utilizando un método denominado pareamiento por puntajes de propensión (*propensity scrore-matching*) (Rosenbaum y Rubin, 1983). Con este enfoque, ya no se requiere que se intente parear a cada unidad inscrita con una unidad no inscrita que tenga exactamente el mismo valor para todas las características de control observables. En cambio, para cada unidad del grupo de tratamiento y del conjunto de no inscritos, se computa la *probabilidad* de que esta unidad se inscriba en el programa (el denominado *puntaje de propensión*) sobre la base de los valores observados de sus características (las variables explicativas). Esta puntuación es un número real entre 0 y 1 que resume la influencia de todas las características observables en la probabilidad de inscribirse en el programa. Deberían utilizarse solo las características observables en la *línea de base* para calcular el puntaje de propensión. Esto se debe a que las características post tratamiento pueden haberse visto afectadas por el propio programa, y el uso de dichas características para identificar a un grupo de comparación pareado sesgaría los resultados. Cuando el tratamiento influye en las características del individuo y se usan aquellas características para parear, se escoge un grupo de comparación que se parece al grupo de tratamiento debido al propio tratamiento. Sin el tratamiento, esas características tendrían un aspecto muy diferente. Esto incumple el requisito básico de una buena estimación del contrafactual, a saber: que el grupo de comparación debe ser similar en todos los aspectos, excepto en el hecho de que el grupo de tratamiento recibe el tratamiento y el grupo de comparación no lo recibe.

Una vez que se ha computado el puntaje de propensión de todas las unidades, aquellas del grupo de tratamiento pueden parearse con unidades en el conjunto de no inscritos que tienen los puntajes de propensión más cercanos.[1] Estas unidades próximas se convierten en el grupo de comparación y se utilizan para producir una estimación del contrafactual. El método de pareamiento por puntajes de propensión intenta imitar la asignación aleatoria a los grupos de tratamiento y comparación escogiendo para el grupo de comparación aquellas unidades que tienen propensiones similares a las unidades del grupo de tratamiento. Dado que el pareamiento de puntajes de propensión no es un método de asignación aleatoria pero intenta imitarlo, pertenece a la categoría de métodos cuasi-experimentales.

La diferencia promedio en los resultados entre las unidades de tratamiento, o inscritas, y sus unidades de comparación correspondientes genera la estimación del impacto del programa. En resumen, el impacto del programa se estima comparando los resultados promedio de un grupo de

tratamiento, o inscrito, y el resultado promedio del subgrupo de unidades estadísticamente pareadas, donde el pareamiento se basa en características observables en los datos disponibles.

Para que el pareamiento por puntajes de propensión produzca estimaciones del impacto de un programa para todas las observaciones tratadas, cada unidad de tratamiento o inscrita debe parearse con una unidad no inscrita.[2] Sin embargo, en la práctica puede ocurrir que, para algunas unidades inscritas, no haya unidades en el conjunto de no inscritos que tengan puntajes de propensión similares. En términos técnicos, puede que se produzca una *falta de rango común*, o falta de superposición, entre los puntajes de propensión del grupo de tratamiento o inscrito y los del conjunto de no inscritos.

El gráfico 8.2 representa un ejemplo de la falta de rango común. En primer lugar, se estima la probabilidad de que cada unidad de la muestra se inscriba en el programa a partir de las características observables de esa unidad, es decir, el puntaje de propensión. El gráfico muestra la distribución de los puntajes de propensión por separado para los inscritos y no inscritos. El problema es que estas distribuciones no se superponen perfectamente. En el medio de la distribución, es relativamente fácil encontrar las parejas porque hay tanto inscritos como no inscritos con estos niveles de puntajes de propensión. Sin embargo, los inscritos con puntajes de propensión cercanos a 1 no se pueden parear con ningún no inscrito porque no hay no inscritos con puntajes de propensión tan altos. Hay tan poca similitud entre las unidades que tienen muchas probabilidades de inscribirse en el programa y

Gráfico 8.2 Pareamiento por puntajes de propensión y rango común

las unidades no inscritas que no se puede encontrar una buena pareja para ellas. De la misma manera, los no inscritos con puntajes de propensión cercanos a 0 no pueden parearse con ningún inscrito porque no hay inscritos que tengan puntajes de propensión tan bajos. Por lo tanto, en los extremos, o colas, de la distribución del puntaje de propensión aparece una falta de rango común. En este caso, el procedimiento de pareamiento estima el efecto local promedio del tratamiento (LATE, por sus siglas en inglés) para las observaciones sobre el rango común.

Los pasos que hay que seguir cuando se aplica un pareamiento por puntajes de propensión se resumen en Jalan y Ravaillon (2003).[3] Primero, se necesitarán encuestas representativas y altamente comparables en las que se puedan identificar las unidades que se inscribieron en el programa y las que no lo hicieron. Segundo, se reúnen las dos muestras y se estima la probabilidad de que cada individuo se inscriba en el programa, a partir de las características individuales observables en la encuesta. Este paso produce el puntaje de propensión. Tercero, se limita la muestra a unidades para las que aparece un rango común en la distribución del puntaje de propensión. Cuarto, para cada unidad inscrita, se identifica un subgrupo de unidades con puntajes de propensión similares. Quinto, se comparan los resultados de las unidades de tratamiento, o inscritas, y las parejas de las unidades de comparación, o no inscritas. La diferencia de los resultados promedio de estos dos subgrupos es la medida del impacto que se puede atribuir al programa para esa observación específica tratada. Sexto, la media de estos impactos individuales arroja una estimación del efecto local promedio del tratamiento. En la práctica, los programas estadísticos habitualmente usados incluyen comandos que realizan los pasos 2 a 6 de manera automática.

En general, es importante recordar tres cuestiones esenciales acerca del pareamiento. En primer lugar, los métodos de pareamiento solo pueden utilizar características *observables* para construir grupos de comparación, dado que las características no observables no se pueden considerar. Si hay alguna característica no observable que influye en la inscripción o no inscripción de la unidad en el programa, y que también influye en el resultado, las estimaciones de impacto obtenidas con el grupo de comparación pareado estarían sesgadas. Para que el resultado del pareamiento no esté sesgado, requiere un supuesto de mucho peso, a saber: que no hay diferencias no observables en el grupo de tratamiento y de comparación que también estén asociadas con los resultados de interés.

Segundo, el pareamiento debe realizarse utilizando solo características que no estén afectadas por el programa. La mayoría de las características que se miden después del comienzo del programa no pertenecerían a esta categoría. Si los datos de línea de base (antes de la intervención) no están disponibles y los únicos datos son los existentes después de que la

intervención ha comenzado, las únicas características que se podrán utilizar para construir una muestra pareada serán aquellas (normalmente pocas) características que no se ven afectadas por un programa, como la edad y el sexo. Aunque se quisiera parear utilizando un conjunto mucho más rico de características, entre ellas los resultados de interés, no se podrá hacerlo porque aquellas están potencialmente afectadas por la intervención. No se recomienda el pareamiento basado únicamente en características posteriores a la intervención. Si hay datos de línea de base disponibles, se puede realizar el pareamiento sobre la base de un conjunto más rico de características, entre ellas, los resultados de interés. Dado que los datos se recopilan antes de la intervención, el programa no puede haber afectado aquellas variables anteriores a la misma. Sin embargo, si hay datos de línea de base sobre los resultados disponibles, no se debería utilizar el método de pareamiento solo, sino que habría que combinarlo con diferencias en diferencias para reducir el riesgo de sesgo. Este procedimiento se detallará en la próxima sección.

Tercero, los resultados de la estimación del método de pareamiento solo son tan buenos como las características que se utilizan para el pareamiento. Si bien es importante poder parear utilizando un gran número de características, lo es aún más poder parear sobre la base de características que determinan la inscripción. Cuanto más se comprenda acerca de los criterios utilizados para la selección de los participantes, en mejores condiciones se estará de construir el grupo de comparación.

La combinación del pareamiento con otros métodos

Aunque la técnica de pareamiento requiere un volumen importante de datos y tiene un riesgo significativo de sesgo, ha sido utilizada para evaluar programas de desarrollo en una amplia gama de contextos. Los usos más convincentes del pareamiento son aquellos que combinan el pareamiento con otros métodos y aquellos que utilizan el método de control sintético. En esta sección, se analizarán las diferencias en diferencias pareadas y el método de control sintético.

Diferencias en diferencias pareadas

Cuando dispone de datos de línea de base sobre los resultados, el pareamiento se puede combinar con diferencias en diferencias para reducir el riesgo de sesgo en la estimación. Como se ha analizado, el simple pareamiento con puntajes de propensión no puede dar cuenta de características no observables que podrían explicar por qué un grupo decide inscribirse en un

programa, y eso también podría afectar los resultados. El pareamiento combinado con diferencias en diferencias al menos tiene en cuenta cualquier característica no observable que sea constante a lo largo del tiempo entre ambos grupos. Se implementa de la siguiente manera:

1. El pareamiento debe realizarse a partir de características observables de la línea de base (como se ha señalado).
2. Para cada unidad inscrita, se debe calcular el cambio en los resultados entre los períodos antes y después (primera diferencia).
3. Para cada unidad inscrita, calcúlese el cambio en los resultados entre los períodos antes y después para la comparación pareada de esta unidad (segunda diferencia).
4. Réstese la segunda diferencia de la primera diferencia, es decir, aplíquese el método de diferencias en diferencias.
5. Por último, calcúlese un promedio de esas dobles diferencias.

Los recuadros 8.1 y 8.2 proporcionan ejemplos de evaluaciones que utilizaron el método de diferencias en diferencias pareadas en la práctica.

Recuadro 8.1: Diferencias en diferencias pareadas: caminos rurales y desarrollo del mercado local en Vietnam

En Vietnam, Mu y Van de Walle (2011) usaron el pareamiento de puntajes de propensión en combinación con el método de diferencias en diferencias para estimar el impacto de un programa de caminos rurales en el desarrollo del mercado local a nivel de la comuna. Entre 1997 y 2001, el gobierno vietnamita rehabilitó 5.000 km de caminos rurales. Los caminos fueron seleccionados según criterios de costo y de densidad demográfica.

Dado que las comunas que se beneficiaron de la reparación de caminos no fueron seleccionadas de forma aleatoria, los investigadores utilizaron el pareamiento de puntajes de propensión para construir un grupo de comparación. Utilizando datos de una encuesta de línea de base, encontraron una diversidad de factores a nivel de la comuna que determinaba si un camino en ella era seleccionado para el programa, como el tamaño de la población, los porcentajes de las minorías étnicas, la calidad de vida, la densidad de los caminos existentes y la presencia de transporte de pasajeros. Estimaron los puntajes de propensión basándose en estas características y limitaron el tamaño de la muestra a la zona de rango común. Esto dio como resultado 94 comunas de tratamiento y 95 comunas de comparación. Para limitar aún más el sesgo de selección

Continúa en la página siguiente.

Recuadro 8.1: Diferencias en diferencias pareadas: caminos rurales y desarrollo del mercado local en Vietnam *(continúa)*

potencial, los investigadores utilizaron diferencias en diferencias para estimar el cambio en las condiciones del mercado local.

Dos años después de iniciado el programa, los resultados indicaron que la rehabilitación de caminos produjo impactos positivos significativos en la presencia y la frecuencia de los mercados locales y en la disponibilidad de servicios. En las comunas de tratamiento se desarrolló un 10% más de nuevos mercados que en las comunas de comparación. En las primeras era más habitual que los hogares cambiaran de actividades agrícolas a actividades más relacionadas con los servicios, como la confección de ropa y las peluquerías. Sin embargo, los resultados variaban considerablemente entre las comunas. En las más pobres, los impactos tendían a ser mayores debido a los menores niveles de desarrollo inicial del mercado. Los investigadores llegaron a la conclusión de que los proyectos de mejora de caminos pequeños pueden tener impactos más importantes si se focalizan en zonas con un desarrollo de mercado inicialmente bajo.

Fuente: Mu y Van de Walle (2011).

Recuadro 8.2: Pareamiento de diferencias en diferencias: suelos de cemento, salud infantil y felicidad de las madres en México

El Programa Piso Firme de México ofrece a los hogares con suelos de tierra hasta 50 m2 de piso de cemento (véase el recuadro 2.1). Piso Firme comenzó como un programa local en el estado de Coahuila, pero fue adoptado a nivel nacional. Cattaneo et al. (2009) aprovecharon la variación geográfica para evaluar el impacto de esta iniciativa para la mejora de la vivienda a gran escala en la salud y en los resultados del bienestar.

Los investigadores utilizaron el método de diferencias en diferencias junto con el pareamiento para comparar a los hogares de Coahuila con familias similares en el estado vecino de Durango, que en el momento de la encuesta todavía no había implementado el programa. Para mejorar la comparabilidad entre los grupos de tratamiento y comparación, los investigadores limitaron su muestra a hogares de las ciudades vecinas situadas justo a ambos lados de la frontera entre los dos estados. En esta muestra, utilizaron técnicas de pareamiento para seleccionar los grupos de tratamiento y comparación que eran más similares. Las características previas al tratamiento que se usaron fueron el porcentaje de hogares con suelos de tierra, el número de hijos pequeños y el número de hogares en cada bloque.

Además del pareamiento, los autores utilizaron variables instrumentales para obtener estimaciones del LATE a partir de

Continúa en la página siguiente.

Recuadro 8.2: Pareamiento de diferencias en diferencias: suelos de cemento, salud infantil y felicidad de las madres en México *(continúa)*

las estimaciones del efecto de la intención de tratar. Con la oferta de un piso de cemento como variable instrumental para determinar si los hogares tenían realmente suelos de cemento encontraron que el programa producía una reducción del 18,2% de la presencia de parásitos, una disminución del 12,4% de la prevalencia de casos de diarrea y una baja del 19,4% de la prevalencia de anemia. Además, pudieron utilizar la variabilidad en el total del espacio del suelo realmente recubierto de cemento para predecir que una completa sustitución de los pisos de tierra por pisos de cemento en un hogar produciría una reducción del 78% de las infecciones parasitarias, una disminución del 59% de los casos de diarrea, una reducción del 81% de la anemia y una mejora del 36% al 96% en el desarrollo cognitivo de los niños. Los autores también recopilaron datos sobre el bienestar de los adultos y llegaron a la conclusión de que los pisos de cemento hacen más felices a las madres, que declararon un aumento del 59% en la satisfacción con la vivienda, un incremento del 69% en la satisfacción con la calidad de vida, una reducción del 52% en la escala de evaluación de la depresión y una disminución del 35% en una escala de evaluación del estrés percibido.

Cattaneo et al. (2009) concluyeron que Piso Firme tiene un impacto absoluto mayor en el desarrollo cognitivo infantil con un costo menor que el programa de transferencias condicionadas de efectivo a gran escala de México, Progresa-Oportunidades, y que otros programas comparables de suplementos nutricionales y estimulación cognitiva temprana. Los pisos de cemento también evitan mejor la proliferación de infecciones parasitarias que el tratamiento habitual de eliminación de parásitos. Los autores afirman que los programas para reemplazar los suelos de tierra con suelos de cemento tienen probabilidades de mejorar la salud de los niños de manera costo-efectiva en contextos similares.

Fuente: Cattaneo et al. (2009).

El método de control sintético

El método de control sintético permite utilizar la estimación del impacto en contextos donde una única unidad (como un país, una empresa o un hospital) es objeto de una intervención o se expone a un evento. En lugar de comparar esta unidad tratada con un grupo de unidades no tratadas, el método usa información sobre las características de la unidad tratada y las unidades no tratadas para construir una unidad de comparación "sintética" o artificial, ponderando cada unidad no tratada de tal manera que la unidad de comparación sintética se asemeje lo más posible a la unidad tratada. Esto requiere una extensa serie de observaciones de las características de la unidad tratada y de las unidades no tratadas a lo

largo del tiempo. Esta combinación de unidades de comparación en una unidad sintética proporciona una mejor comparación para la unidad tratada que cualquier unidad no tratada individualmente. El recuadro 8.3 presenta un ejemplo de una evaluación para la cual se empleó el método de control sintético.

Recuadro 8.3: El método de control sintético: los efectos económicos de un conflicto terrorista en España

Abadie y Gardeazábal (2003) utilizaron el método de control sintético para investigar los efectos económicos de un conflicto terrorista en el País Vasco. A comienzos de los años setenta el País Vasco era una de las regiones más ricas de España. Sin embargo, hacia finales de los años noventa, después de 30 años de conflicto, había caído hasta la sexta posición en el producto interno bruto (PIB) per cápita. En los albores de los atentados terroristas a comienzos de los años setenta, el País Vasco era diferente de otras regiones de España en características que, según se piensa, están relacionadas con el potencial de crecimiento económico. Por lo tanto, la comparación entre el crecimiento del PIB de la economía vasca y del resto de España reflejaría tanto el efecto del terrorismo como el efecto de estas diferencias en los factores determinantes del crecimiento económico antes del comienzo del terrorismo. En otras palabras, el enfoque de diferencias en diferencias produciría resultados sesgados del impacto del terrorismo en el crecimiento económico del País Vasco. Para lidiar con esta situación, los autores utilizaron una combinación de otras regiones españolas, de modo de construir una región de comparación "sintética".

Fuente: Abadie y Gardeazábal (2003).

Evaluación del impacto del HISP: la utilización de técnicas de pareamiento

Después de conocer las técnicas de pareamiento, puede que uno se pregunte si podría usarlas para estimar el impacto del Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés). Por ejemplo, se decide utilizar técnicas de pareamiento para seleccionar un grupo de hogares no inscritos que parecen similares a los hogares inscritos a partir de las características observables de la línea de base. Para hacer esto, se utiliza el paquete de pareamiento del *software* estadístico. Primero, se debe estimar la probabilidad de que un hogar se inscriba en el programa

sobre la base de los valores observados de las características (las variables explicativas), como la edad del jefe de hogar y del cónyuge, su nivel de estudios, si el jefe del hogar es mujer, si el hogar es indígena, etc.

A continuación, se llevará a cabo un pareamiento considerando dos escenarios. En el primero, hay un gran conjunto de variables para predecir la inscripción, entre ellas las características socioeconómicas del hogar. En el segundo, hay escasa información para predecir la inscripción (solo el nivel de estudios y la edad del jefe de hogar). Como se muestra en el cuadro 8.1, la probabilidad de que un hogar se inscriba en el programa es menor si el jefe de hogar es mayor, si tiene más estudios, si es mujer, o si el hogar cuenta con baño o posee un terreno más grande. Por el contrario, ser indígena, tener más miembros en el hogar, tener un suelo de tierra y vivir más lejos de un hospital son factores que

Cuadro 8.1 Estimación del puntaje de propensión a partir de características observables de la línea de base

Variable dependiente: inscritos = 1	**Todo el conjunto de variables explicativas**	**Conjunto limitado de variables explicativas**
Variables explicativas: características observables en la línea de base	**Coeficiente**	**Coeficiente**
Edad del jefe del hogar (años)	–0,013**	–0,021**
Edad del cónyuge (años)	–0,008**	–0,041**
Nivel educativo del jefe del hogar (años)	–0,022**	
Nivel educativo del cónyuge (años)	–0,016*	
Jefe del hogar es mujer =1	–0,020	
Indígena = 1	0,161**	
Número de miembros del hogar	0,119**	
Suelo de tierra = 1	0,376**	
Baño = 1	–0,124**	
Hectáreas de terreno	–0,028**	
Distancia del hospital (km)	0,002**	
Constante	–0,497**	0,554**

Nota: Regresión *probit*. La variable dependiente equivale a 1 si el hogar está inscrito en el HISP y 0 en caso contrario. Los coeficientes representan la contribución de cada variable explicativa a la probabilidad de que un hogar se inscriba en el HISP.
Nivel de significancia: * = 5%; ** = 1%.

aumentan la probabilidad de que un hogar se inscriba en el programa. Por lo tanto, en general, parecería que los hogares más pobres y con menor nivel educativo tienen más probabilidades de inscribirse, lo cual es una buena noticia para un programa que se focaliza en las personas pobres.

Ahora que el *software* ha estimado la probabilidad de que todos los hogares se inscriban en el programa (el puntaje de propensión), se verifica la distribución del puntaje de propensión para los hogares de comparación inscritos y pareados. El gráfico 8.3 muestra que el rango común (cuando se utiliza todo el conjunto de variables explicativas) se extiende por toda la distribución del puntaje de propensión. De hecho, ninguno de los hogares inscritos queda marginado de la zona de rango común. En otras palabras, se puede encontrar un hogar de comparación como pareja para cada uno de los hogares inscritos.

Gráfico 8.3 Pareamiento para el HISP: rango común

Se decide utilizar el pareamiento de vecino más próximo, es decir, se le pide al *software* que, para cada hogar inscrito, encuentre el hogar no inscrito que tiene el puntaje de propensión más cercano al hogar inscrito. El *software* limita la muestra a aquellos hogares en los grupos de inscritos y no inscritos para los que puede encontrar una pareja en el otro grupo.

Para obtener la estimación de impacto utilizando el método de pareamiento, primero se calcula el impacto para cada hogar inscrito individualmente (utilizando el hogar de comparación pareado de cada hogar) y luego se calcula el promedio de esos impactos individuales. El cuadro 8.2 muestra que el impacto estimado al aplicar este procedimiento es una reducción de US$9,95 en los gastos en salud de los hogares.

Por último, el *software* también permite computar el error estándar en la estimación de impacto utilizando la regresión lineal (cuadro 8.3).[4]

Cuadro 8.2 Evaluación del HISP: pareamiento de las características de línea de base y comparación de medias

	Inscritos	Comparación pareada	Diferencia
Gasto en salud de los hogares (US$)	7,84	17,79 (usando el conjunto de variables explicativas)	–9,95
		19,9 (utilizando un conjunto limitado de variables explicativas)	–11,35

Nota: Este cuadro compara los gastos en salud promedio de los hogares en los hogares inscritos y las parejas de hogares de comparación.

Cuadro 8.3 Evaluación del HISP: pareamiento de las características de línea de base y análisis de regresión

	Regresión lineal (pareamiento en todo el conjunto de variables explicativas)	Regresión lineal (pareamiento en conjunto limitado de variables explicativas)
Impacto estimado en los gastos en salud de los hogares (US$)	–9,95** (0,24)	–11,35** (0,22)

Nota: Los errores estándar están entre paréntesis.
Nivel de significancia: ** = 1%.

Se observa también que en los datos de las encuestas se cuenta con información de los resultados de línea de base, de modo que se deciden utilizar las diferencias en diferencias pareadas además de usar todo el conjunto de variables explicativas. Es decir, se calcula la diferencia en los gastos en salud de los hogares en el seguimiento entre hogares inscritos y

hogares de comparación pareados; se computa la diferencia en los gastos en salud de los hogares en la línea de base entre los hogares inscritos y las parejas de comparación; y luego se calcula la diferencia entre estas dos diferencias. El cuadro 8.4 muestra el resultado de este enfoque de diferencias en diferencias pareadas.

Cuadro 8.4 Evaluación del HISP: método de diferencias en diferencias combinado con pareamiento en las características de línea de base

		Inscritos	Comparaciones pareadas utilizando el conjunto de variables explicativas	Diferencia
Gastos en salud de los hogares (US$)	Seguimiento	7,84	17,79	–9,95
	Línea de base	14,49	15,03	0,54
				Diferencias en diferencias pareadas =–9,41** (0,19)

Nota: Los errores estándar están entre paréntesis y el cálculo se realizó utilizando una regresión lineal.

Nivel de significancia: ** = 1%.

Pregunta HISP 7

A. ¿Cuáles son los supuestos básicos necesarios para aceptar estos resultados sobre la base del método de pareamiento?

B. ¿Por qué los resultados del método de pareamiento son diferentes si se utiliza todo el conjunto vs. el conjunto limitado de variables explicativas?

C. ¿Qué sucede cuando se compara el resultado del método de pareamiento con el resultado de la asignación aleatoria? ¿Por qué los resultados son tan diferentes en el pareamiento con un conjunto limitado de variables explicativas? ¿Por qué el resultado es más parecido cuando se realiza el pareamiento en todo el conjunto de variables explicativas?

D. A partir del resultado del método de pareamiento, ¿debería ampliarse el HISP a escala nacional?

Limitaciones del método de pareamiento

Aunque los procedimientos de pareamiento se pueden aplicar en numerosos contextos, independientemente de las reglas de asignación de un programa, tienen varias limitaciones importantes. En primer lugar, requieren conjuntos de datos amplios sobre grandes muestras de unidades, e incluso cuando estos están disponibles, puede que se produzca una falta de rango común entre el grupo de tratamiento, o inscrito, y el conjunto de no participantes. En segundo lugar, solo se puede aplicar el pareamiento basándose en características observables; por definición, no se pueden incorporar las características no observables en el cálculo del puntaje de propensión. Por lo tanto, para que el procedimiento de pareamiento identifique un grupo de comparación válido, no deben existir diferencias sistemáticas en las características no observables entre las unidades de tratamiento y las unidades de comparación pareadas[5] que podrían influir en el resultado (Y). Dado que no se puede *demostrar* que existen esas características no observables que influyen en la participación y en los resultados, se debe *suponer* que no existen. Normalmente se trata de un supuesto de mucho peso. A pesar de que el pareamiento contribuye a controlar por características básicas *observables*, nunca se puede descartar el sesgo que nace de las características *no observables*. En resumen, el supuesto de que no se ha producido un sesgo de selección debido a las características no observables es de mucho peso y, lo que es más problemático, no puede comprobarse.

El pareamiento por sí solo suele ser menos robusto que los otros métodos de evaluación analizados, dado que requiere el fuerte supuesto de que no hay características no observables que influyan simultáneamente en la participación en el programa y en sus resultados. Por otro lado, la asignación aleatoria, la variable instrumental y el diseño de regresión discontinua no requieren el supuesto indemostrable de que no hay tales variables no observables. Tampoco requieren muestras tan grandes o características básicas tan amplias como el pareamiento por puntajes de propensión.

En la práctica, los métodos de pareamiento suelen usarse cuando no es posible recurrir a las opciones de asignación aleatoria, variable instrumental y diseño de regresión discontinua. El denominado *pareamiento ex post* es muy riesgoso cuando no hay datos de línea de base disponibles sobre el resultado de interés o de las características básicas. Si una evaluación utiliza datos de encuestas que fueron recopilados después del comienzo del programa (es decir, ex post) para deducir las características básicas de las

unidades de la línea de base y luego emparejar el grupo tratado con un grupo de comparación empleando esas características deducidas, puede emparejar involuntariamente basándose en características que también fueron afectadas por el programa; en ese caso, el resultado de estimación sería inválido o estaría sesgado.

Por el contrario, cuando se dispone de datos de línea de base, el pareamiento basado en las características básicas puede ser muy útil si se combina con otras técnicas, como el método de diferencias en diferencias, lo que permite corregir por las diferencias entre los grupos que son fijas a lo largo del tiempo. El pareamiento también es más fiable cuando se conocen las reglas de asignación del programa y las variables fundamentales, en cuyo caso el pareamiento se puede llevar a cabo con esas variables.

A estas alturas, es probable que quede claro que las evaluaciones de impacto se diseñan mejor antes de que un programa comience a ser implementado. Una vez que el programa ha comenzado, si hay que influir en cómo se asigna y no se han recopilado datos de línea de base, habrá pocas o ninguna opción rigurosa para la evaluación de impacto.

Verificación: el pareamiento

El pareamiento se basa en el supuesto de que las unidades inscritas y no inscritas son similares en términos de cualquier variable no observable que podría influir tanto en la probabilidad de participar en el programa como en el resultado.

- ✓ ¿La participación en el programa está determinada por variables que no se pueden observar? Esto no se puede comprobar directamente, de modo que para orientarse habrá que fiarse de la teoría, del sentido común y del conocimiento adecuado del contexto de la evaluación de impacto.
- ✓ ¿Las características observables están bien equilibradas entre los subgrupos pareados? Compárense las características observables de cada grupo de tratamiento y su grupo de unidades de comparación pareados en la línea de base.
- ✓ ¿Se puede encontrar una unidad de comparación pareada para cada unidad de tratamiento? Verifíquese si hay un rango común suficiente en la distribución de los puntajes de propensión. Las pequeñas zonas de rango común o superposición señalan que las personas inscritas y no inscritas son muy diferentes, y aquello arroja dudas sobre si el pareamiento es un método creíble.

Otros recursos

- Para material de apoyo relacionado con este libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para más información, consúltese P. Rosenbaum (2002), *Observational Studies* (2da. edición), Springer Series in Statistics. Nueva York: Springer-Verlag.
- Para más información sobre la implementación del pareamiento por puntajes de propensión, véase C. Heinrich, A. Maffioli y G. Vásquez (2010), "A Primer for Applying Propensity-Score Matching. Impact-Evaluation Guidelines." Nota técnica del BID-TN-161. Washington, D.C.: BID.

Notas

1. Nota técnica: en la práctica, se utilizan muchas definiciones de lo que constituye la unidad más próxima o cercana para llevar a cabo un pareamiento. Las unidades de control más cercanas se pueden definir sobre la base de una estratificación del puntaje de propensión –la identificación de los vecinos más próximos de la unidad de tratamiento, considerando la distancia, dentro de un determinado radio– o utilizando técnicas de núcleo. Se considera una buena práctica verificar la robustez de los resultados del pareamiento empleando diversos algoritmos de pareamiento. Para más detalles, véase Rosenbaum (2002).
2. En este libro, el análisis del pareamiento se centra en un pareamiento de uno a uno. No se analizarán otros tipos de pareamiento, como el de uno a varios o el de reemplazo/sin reemplazo. Sin embargo, en todos los casos el marco conceptual descrito aquí seguiría vigente.
3. En Rosenbaum (2002) se puede encontrar un análisis detallado del pareamiento.
4. Nota técnica: cuando las puntuaciones de propensión de las unidades inscritas no están plenamente cubiertas por el área de rango común, los errores estándar deberían estimarse utilizando un muestreo autodocimante en lugar de una regresión lineal.
5. Para los lectores que tienen conocimientos de econometría, esto significa que la participación es independiente de los resultados, dadas las características básicas utilizadas para realizar el pareamiento.

Referencias bibliográficas

Abadie, A. y J. Gardeazábal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 113–32.

Cattaneo, M. D., S. Galiani, P. J. Gertler, S. Martínez y R. Titiunik. 2009. "Housing, Health, and Happiness." *American Economic Journal: Economic Policy* 1 (1): 75–105.

Heinrich, C., A. Maffioli y G. Vázquez. 2010. "A Primer for Applying Propensity-Score Matching. Impact-Evaluation Guidelines." Nota técnica del BID-TN-161. Washington, D.C.: BID.

Jalan, J. y M. Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching." *Journal of Business & Economic Statistics* 21 (1): 19–30.

Mu, R. y D. Van de Walle. 2011. "Rural Roads and Local Market Development in Vietnam." *Journal of Development Studies* 47 (5): 709–34.

Rosenbaum, P. 2002. *Observational Studies* (2da. edición), Springer Series in Statistics. Nueva York: Springer-Verlag.

Rosenbaum, P. y D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies of Causal Effects." *Biometrika* 70 (1): 41–55.

CAPÍTULO 9

Cómo abordar las dificultades metodológicas

Efectos heterogéneos del tratamiento

Ya se ha visto que la mayoría de los métodos de evaluación de impacto solo producen estimaciones válidas del contrafactual bajo supuestos específicos. El principal riesgo de cualquier método es que sus supuestos fundamentales no sean válidos, lo que genera estimaciones sesgadas del impacto del programa que se evalúa. Sin embargo, también hay otros riesgos comunes en la mayoría de las metodologías que se han analizado. En este capítulo, se examinarán los principales.

Un tipo de riesgo surge cuando se estima el impacto de un programa en todo un grupo y los resultados ocultan algunas diferencias en las respuestas al tratamiento de los diferentes receptores, es decir, los efectos heterogéneos del tratamiento. La mayoría de los métodos de evaluación de impacto supone que un programa influye en los resultados de una manera simple y lineal para todas las unidades de la población.

Sin embargo, si se piensa que diferentes subpoblaciones pueden haber vivido el impacto de un programa de manera muy diferente, puede que convenga tener muestras estratificadas para cada subpoblación. Supóngase, por ejemplo, que a uno le interesa conocer el impacto de un programa de comida escolar en las niñas, pero solo el 10% del alumnado está conformado por niñas. En ese caso, puede que incluso una muestra aleatoria

"grande" de alumnos no contenga un número suficiente de niñas como para estimar el impacto del programa en ellas. Para el diseño de la muestra de evaluación, convendría estratificar la misma basándose en el sexo, e incluir un número suficientemente grande de niñas a fin de poder detectar un determinado tamaño del efecto.

Efectos no intencionados en la conducta

Cuando se lleva a cabo una evaluación de impacto, también puede suceder que se induzca a respuestas no intencionadas en la conducta de la población que se estudia, a saber:

- El *efecto Hawthorne* ocurre cuando debido al mero hecho de saber que están siendo observadas, las unidades se comportan de manera diferente (véase el recuadro 9.1).

Recuadro 9.1: Cuentos tradicionales de la evaluación de impacto: el efecto Hawthorne y el efecto John Henry

La expresión *efecto Hawthorne* se refiere a los experimentos llevados a cabo entre 1924 y 1932 en el Hawthorne Works, una fábrica de equipos eléctricos en el estado de Illinois. Los experimentos probaron el impacto del cambio en las condiciones de trabajo (como aumentar o disminuir la intensidad de la luz) en la productividad de los trabajadores, y llegaron a la conclusión de que cualquier cambio en las condiciones de trabajo (más o menos luz, más o menos períodos de descanso, etc.) producía un aumento de la productividad. Esto se interpretó como un efecto de observación, es decir, los trabajadores que formaban parte del experimento se vieron a sí mismos como algo especial y su productividad aumentó debido a esto, y no debido al cambio en las condiciones de trabajo. Si bien los experimentos originales posteriormente fueron objeto de polémicas y en alguna medida se los desacreditó, la expresión efecto Hawthorne permaneció.

En cuanto al *efecto John Henry*, la expresión fue acuñada por Gary Saretsky en 1972 para referirse al legendario héroe popular John Henry, el "hombre del taladro de acero" encargado de horadar las rocas con un taladro de acero para preparar los agujeros de los explosivos durante la construcción de un túnel de ferrocarril. Según cuenta la leyenda, cuando Henry supo que se le comparaba con un taladro de acero, trabajó esforzándose mucho más para superar a la propia máquina. Desafortunadamente, falleció como consecuencia de ello. Sin embargo, la expresión sigue vigente para describir cómo las unidades de comparación a veces se esfuerzan más para compensar el hecho de no ser objeto de un tratamiento.

Fuentes: Landsberger (1958).

- El *efecto John Henry* se produce cuando las unidades de comparación se esfuerzan más para compensar el hecho de no ser objeto del tratamiento (véase el recuadro 9.1).
- La *anticipación* puede generar otro tipo de efecto no intencionado en la conducta. En una aleatorización por fases, puede que las unidades del grupo de comparación esperen recibir el programa en el futuro y comiencen a cambiar su comportamiento antes de que el programa realmente se materialice.
- El *sesgo por sustitución* es otro efecto en la conducta que influye en el grupo de comparación: las unidades que no fueron seleccionadas para ser objeto del programa pueden encontrar buenos sustitutos gracias a su propia iniciativa.

Las respuestas en la conducta que afectan de manera desproporcionada al grupo de comparación constituyen un problema porque pueden socavar la validez interna de los resultados de la evaluación, aunque se use la asignación aleatoria como método de evaluación. Un grupo de comparación que se esfuerza más para compensar el hecho de no ser objeto de un tratamiento, o que cambia su conducta en previsión del programa, no es una buena representación del contrafactual.

Si se tiene algún motivo para creer que se pueden producir estas respuestas no intencionadas en la conducta, a veces una opción es constituir otros grupos de comparación que no se vean en absoluto afectados por la intervención, es decir, un grupo que permita explícitamente comprobar dichas respuestas. También puede que sea una buena idea recopilar datos cualitativos con el fin de entender mejor las respuestas en la conducta.

Imperfección del cumplimiento

La *imperfección del cumplimiento* es la discrepancia entre la condición asignada del tratamiento y la condición real del mismo. La imperfección del cumplimiento se produce cuando algunas unidades asignadas al grupo de tratamiento no reciben tratamiento, y cuando algunas unidades asignadas al grupo de comparación reciben tratamiento. En el capítulo 5 se estudia la imperfección del cumplimiento en referencia a la asignación aleatoria, si bien la imperfección del cumplimiento también se puede producir con el diseño de regresión discontinua (como se señala en el capítulo 6) y con diferencias en diferencias (capítulo 7). Antes de que se puedan interpretar las estimaciones de impacto que genera cualquier método, es necesario saber si se ha producido una imperfección del cumplimiento en el programa.

La imperfección del cumplimiento puede tener lugar de diversas maneras:

- No todos los participantes previstos participan realmente en el programa. A veces, algunas unidades asignadas a un programa deciden no participar.
- Algunos participantes previstos son excluidos del programa debido a errores administrativos o de ejecución.
- Se ofrece el programa por error a algunas unidades del grupo de comparación, que se inscriben en él.
- Algunas unidades del grupo de comparación consiguen participar en el programa a pesar de que no se les ofrece.
- El programa se asigna a partir del índice continuo de elegibilidad, pero no se aplica estrictamente el umbral de elegibilidad.
- Se produce una *migración selectiva* en función de la condición del tratamiento. Por ejemplo, puede que la evaluación compare los resultados en los municipios tratados y no tratados, pero las personas pueden decidir trasladarse a otro municipio si no les agrada la condición de tratamiento de su municipio.

En general, ante una situación de imperfección del cumplimiento, los métodos normales de evaluación de impacto producen estimaciones de la intención de tratar. Sin embargo, se pueden obtener estimaciones del tratamiento en los tratados a partir de las estimaciones de la intención de tratar mediante el método de variables instrumentales.

En el capítulo 5 se explicó la idea básica para lidiar con la imperfección del cumplimiento en el contexto de la asignación aleatoria. Mediante un ajuste del porcentaje de cumplidores en la muestra de la evaluación, se pudo recuperar el efecto local promedio del tratamiento en los cumplidores a partir de la estimación de la intención de tratar. Este "ajuste" puede ampliarse a otros métodos mediante la aplicación del enfoque más general de variables instrumentales. La variable instrumental contiene una fuente externa de variación que ayuda a eliminar o corregir el sesgo que puede derivarse de la imperfección en el cumplimiento. En el caso de la asignación aleatoria con imperfección en el cumplimiento, se utiliza una variable 0/1 (denominada *dummy*) que asume el valor de 1 si la unidad estaba asignada originalmente al grupo de tratamiento, y de 0 si la unidad estaba originalmente asignada al grupo de comparación. Durante la fase de análisis, la variable instrumental se usa con frecuencia en el contexto de una *regresión en dos fases* que permite identificar el impacto del tratamiento en los cumplidores.

La lógica del método de variable instrumental puede aplicarse al contexto de otros métodos de evaluación:

- En el contexto del diseño de regresión discontinua, debe utilizarse una variable 0/1 que indique si la unidad se encuentra en el lado no elegible o elegible de la puntuación límite.
- En el contexto de la migración selectiva, una posible variable instrumental para la ubicación del individuo después del comienzo del programa sería la ubicación del individuo antes del anuncio del programa.

A pesar de la posibilidad de abordar la imperfección en el cumplimiento utilizando variables instrumentales, es necesario recordar tres puntos:

1. Desde un punto de vista técnico, no es deseable que una gran parte del grupo de comparación se inscriba en el programa. A medida que aumenta la proporción del grupo de comparación que se inscribe en el programa, la fracción de "cumplidores" en la población disminuirá, y el efecto local promedio del tratamiento estimado con el método de variable instrumental será válido solo para una fracción cada vez más pequeña de la población de interés. Si esto se extiende demasiado, puede que los resultados pierdan toda relevancia para las políticas, dado que ya no serían aplicables a una parte suficientemente grande de la población de interés.
2. Tampoco es deseable que una parte grande del grupo de tratamiento siga sin inscribirse. Una vez más, a medida que la fracción del grupo de tratamiento que se inscribe en el programa disminuye, también lo hace la fracción de "cumplidores" de la población. El efecto promedio del tratamiento estimado con el método de variable instrumental será válido solo para una fracción cada vez menor de la población de interés.
3. Como ya se trató en el capítulo 5, el método de variables instrumentales es válido solo en ciertas circunstancias; decididamente no es una solución universal.

El efecto de derrame

Los derrames (o efectos de derrame) son otro problema habitual a los que se enfrentan las evaluaciones, sea que se aplique el método de asignación aleatoria, el de diseño de regresión discontinua o el de diferencias en diferencias. Un *derrame* se produce cuando una intervención afecta a un no

participante, y puede ser positivo o negativo. Hay cuatro tipos de efectos de derrame, según Angelucci y Di Maro (2015):

- *Externalidades.* Se trata de efectos que van de los sujetos tratados a los sujetos no tratados. Por ejemplo, vacunar contra la gripe a los niños de un pueblo reduce la probabilidad de que los habitantes no vacunados del mismo pueblo contraigan esa enfermedad. Se trata de un ejemplo de externalidades positivas. Las externalidades también pueden ser negativas. Por ejemplo, los cultivos de un agricultor pueden verse parcialmente destruidos si su vecino aplica un herbicida en su propio terreno y parte del herbicida cae sobre el otro lado de la línea divisoria de la propiedad.

- *Interacción social.* Los efectos de derrame pueden ser el producto de interacciones sociales y económicas entre poblaciones tratadas y no tratadas, que conducen a impactos indirectos en los no tratados. Por ejemplo, un alumno que recibe una *Tablet* como parte de un programa de mejora del aprendizaje puede compartir el dispositivo con otro alumno que no participa en el programa.

- *Efectos de equilibrio del contexto.* Estos efectos se producen cuando una intervención influye en las normas comportamentales o sociales dentro de un determinado contexto, como una localidad tratada. Por ejemplo, aumentar la cantidad de recursos que reciben los centros de salud tratados de manera que puedan ampliar su gama de servicios puede influir en las expectativas de la población a propósito de cuál debería ser el nivel de los servicios ofrecidos en todos los centros de salud.

- *Efectos de equilibrio general.* Estos efectos se producen cuando las intervenciones influyen en la oferta y demanda de bienes y servicios y, por ende, cambian el precio de mercado de esos servicios. Por ejemplo, un programa que entrega vales a las mujeres pobres para que utilicen los centros privados para dar a luz puede aumentar la demanda de servicios en los centros privados, lo que incrementaría el precio del servicio para todos. El recuadro 9.2 presenta un ejemplo de externalidades negativas debido a efectos de equilibrio general en el contexto de un programa de capacitación laboral.

Si el no participante que experimenta el derrame pertenece al grupo de comparación, el efecto derrame viola el requisito básico de que el resultado de una unidad no debería verse afectado por la asignación concreta de tratamientos a otras comunidades. Este *supuesto de estabilidad del valor de la unidad de tratamiento* (o SUTVA, por sus siglas en inglés, *stable unit treatment value assumption*) es necesario para asegurar que la asignación aleatoria produzca estimaciones no sesgadas del impacto. Si el grupo de control se ve

Recuadro 9.2: Externalidades negativas debidas a efectos de equilibrio general: asistencia para la colocación laboral y resultados del mercado de trabajo en Francia

Los programas de asistencia para la colocación laboral son populares en numerosos países desarrollados. Los gobiernos contratan a una entidad externa para que ayude a los trabajadores desempleados en su búsqueda de empleo. Numerosos estudios llegan a la conclusión de que estos programas de asesoría tienen un impacto significativo y positivo en quienes buscan empleo.

Crépon et al. (2013) investigaron si procurar asistencia laboral a trabajadores jóvenes y con estudios que buscaban empleo en Francia podría tener efectos negativos en otros jóvenes que buscan empleo pero que no tienen el apoyo del programa. Según su hipótesis, podría existir un mecanismo de derrame, es decir, cuando el mercado laboral no está creciendo demasiado, ayudar a una persona que busca empleo a encontrarlo puede producirse en desmedro de otra persona que busca empleo y que, de otra manera, podría haber obtenido el empleo que obtuvo el trabajador asesorado. Para investigar esta hipótesis, llevaron a cabo un experimento aleatorio que incluía 235 mercados laborales (en las ciudades) de Francia. Estos mercados de trabajo fueron asignados aleatoriamente a uno de cinco grupos, que variaban en cuanto al porcentaje de buscadores de empleo que eran asignados al tratamiento de asesoría (0%, 25%, 50%, 75% y 100%). En cada mercado laboral, los buscadores de empleo elegibles eran asignados de forma aleatoria al tratamiento, siguiendo esta proporción. Al cabo de ocho meses, los autores encontraron que los jóvenes desempleados asignados al programa tenían probabilidades significativamente mayores de encontrar un empleo estable que aquellos que no habían sido asignados. No obstante, al parecer esto se produjo parcialmente a expensas de los trabajadores elegibles que no se beneficiaron del programa.

Fuente: Crépon et al. (2013).

indirectamente afectado por el tratamiento recibido por el grupo de tratamiento (por ejemplo, los alumnos del grupo de comparación que usan las *Tablets* de los alumnos del grupo de tratamiento), la comparación no representa con precisión qué habría ocurrido en el grupo de tratamiento en ausencia de tratamiento (el contrafactual).

Si el no participante que experimenta el derrame no pertenece al grupo de comparación, el supuesto SUTVA sería válido y el grupo de comparación seguiría proporcionando una buena estimación del contrafactual. Sin embargo, aún habría que medir el derrame, porque representa un impacto real del programa. En otras palabras, la comparación de los resultados de los grupos de tratamiento y comparación generaría estimaciones no sesgadas del impacto del tratamiento en el grupo tratado, pero esto no tendría en cuenta el impacto del programa en *otros* grupos.

Un ejemplo clásico de efectos de derrame debido a externalidades se presenta en Kremer y Miguel (2004), que analizaron el impacto de administrar una medicación antiparasitaria a niños en escuelas de Kenia (recuadro 9.3). Los parásitos intestinales pueden transmitirse de una persona a otra a través del contacto con materia fecal contaminada. Cuando un niño recibe el remedio antiparasitario, su "carga de parásitos" disminuye, pero también lo hará la carga de parásitos de las personas que viven en su entorno, dado que ya no entrarán en contacto con los parásitos del niño. Por lo tanto, en el ejemplo de Kenia, cuando se administró el remedio a los niños de una escuela, se beneficiaron no solo esos niños (beneficio directo) sino también los niños de las escuelas vecinas (beneficios indirectos).

Como se describe en el gráfico 9.1, la campaña antiparasitaria de las escuelas del grupo A también disminuye el número de parásitos que afectan a los niños que no pertenecen a las escuelas del grupo A. Concretamente, puede reducir la carga de parásitos que afectan a los niños que van a las escuelas del grupo de comparación B, situadas cerca de las escuelas del grupo A. Sin embargo, las escuelas de comparación que se hallan más lejos de las escuelas del grupo A –las llamadas escuelas del grupo C– no experimentan dichos efectos de derrame porque el remedio administrado en el grupo A no elimina los parásitos que afectan a los niños que van a las escuelas del grupo C. La evaluación y sus resultados se estudian con más detalle en el recuadro 9.3.

Recuadro 9.3: Trabajando con los efectos de derrame: remedios antiparasitarios, externalidades y educación en Kenia

El Proyecto de Tratamiento Antiparasitario de las escuelas primarias de Busia, Kenia, fue diseñado para probar diversos aspectos de los tratamientos antiparasitarios y de la prevención. La iniciativa fue un programa de la organización holandesa sin fines de lucro International Child Support Africa, en cooperación con el Ministerio de Salud de Kenia. El proyecto abarcaba 75 escuelas con una matriculación total de más de 30.000 alumnos de 6 a 18 años. Los niños fueron tratados con remedios antiparasitarios de conformidad con las recomendaciones de la Organización Mundial de la Salud (OMS) y también recibieron educación antiparasitaria preventiva con charlas sobre la salud, pósteres y capacitación de los profesores.

Debido a limitaciones administrativas y financieras, el programa se llevó a cabo según el orden alfabético de las escuelas. El primer grupo de 25 escuelas comenzó en 1998, el segundo grupo en 1999 y el tercer grupo en 2001. Mediante una selección aleatoria de las escuelas, Kremer y Miguel

Continúa en la página siguiente.

Recuadro 9.3: Trabajando con los efectos de derrame: remedios antiparasitarios, externalidades y educación en Kenia *(continúa)*

(2004) pudieron estimar el impacto del tratamiento antiparasitario en un establecimiento e identificar los derrames en otras escuelas utilizando una variación exógena de la cercanía entre las escuelas de comparación y las de tratamiento. Aunque el cumplimiento del diseño aleatorio fue relativamente alto (el 75% de los alumnos asignados al grupo de tratamiento recibió los medicamentos antiparasitarios y solo un pequeño porcentaje del grupo de comparación recibió tratamiento), los investigadores pudieron aprovechar el no cumplimiento para determinar las externalidades de salud, o derrames, en las escuelas.

El efecto directo de las intervenciones fue una reducción de las infecciones parasitarias moderadas a graves en 26 puntos porcentuales para los alumnos que tomaban la medicación. Entretanto, las infecciones moderadas a graves entre los alumnos que asistían a las escuelas de tratamiento pero no tomaban la medicación disminuyeron en 12 puntos porcentuales a través de un efecto de derrame indirecto. También se observaron externalidades entre las escuelas.

Dado que el costo del tratamiento antiparasitario es tan bajo y que los efectos en la salud y la educación son relativamente altos, los autores llegaron a la conclusión de que el tratamiento antiparasitario es una manera relativamente costo-efectiva para mejorar las tasas de participación en las escuelas. El estudio también muestra que las enfermedades tropicales como los parásitos pueden desempeñar un importante rol en los resultados educativos, lo cual fortalece los argumentos de que la alta carga de infecciones existente en África puede ser uno de los factores que explica su bajo ingreso. Por lo tanto, Kremer y Miguel sostienen que el estudio es un sólido argumento a favor de las subvenciones públicas a los tratamientos contra las infecciones, con beneficios de derrame similares en los países en desarrollo.

Fuente: Kremer y Miguel (2004).

Diseño de una evaluación de impacto que tiene en cuenta los derrames

Supóngase que se diseña una evaluación de impacto para un programa donde es probable que se produzcan derrames. ¿Cómo se enfocaría esto? Lo primero es entender que el objetivo de la evaluación necesita ser más amplio. Mientras que una evaluación estándar pretende estimar el impacto (o efecto causal) de un programa en un resultado de interés para las unidades que reciben el tratamiento, una evaluación con efectos de derrame tendrá que responder a dos preguntas:

1. *La pregunta estándar sobre la evaluación del impacto directo.* ¿Cuál es el impacto (o efecto causal) de un programa en un resultado de interés para

Gráfico 9.1 Un ejemplo clásico de efecto de derrame: externalidades positivas de la administración de remedios antiparasitarios a los niños de las escuelas

las unidades que reciben el tratamiento? Se trata del impacto directo que el programa tiene en los grupos tratados.

2. *Una segunda pregunta sobre la evaluación del impacto indirecto.* ¿Cuál es el impacto (o efecto causal) de un programa en un resultado de interés en las unidades que *no* reciben el tratamiento? Se trata del impacto indirecto que el programa tiene en los grupos no tratados.

Para estimar el impacto directo en los grupos tratados, habrá que elegir el grupo de comparación de tal manera que no se vea afectado por los derrames. Por ejemplo, puede ponerse como condición que los pueblos, clínicas u hogares de tratamiento y comparación estén situados lo suficientemente lejos unos de otros de manera que los derrames sean poco probables.

Para estimar el impacto indirecto en los grupos no tratados, debería identificarse para cada grupo no tratado un grupo de comparación adicional que

pueda verse afectado por los derrames. Por ejemplo, los trabajadores comunitarios de la salud pueden realizar visitas domiciliarias para proporcionar información a los padres acerca de los beneficios de una dieta variada mejorada para los niños. Supóngase que los trabajadores comunitarios de la salud solo visitan algunos hogares de un pueblo determinado. Uno puede estar interesado en los efectos de derrame sobre los niños de los hogares no visitados, en cuyo caso necesitaría hallar un grupo de comparación para estos niños. Al mismo tiempo, puede ser que la intervención también afecte la variedad de la dieta de los adultos. Si tal efecto indirecto es de interés para la evaluación, se necesitaría también un grupo de comparación para los adultos. A medida que aumente el número de canales potenciales de derrame, el diseño puede complicarse con relativa rapidez.

Las evaluaciones con efectos de derrame plantean ciertos problemas específicos. Por ejemplo, cuando los efectos de derrame son probables, es importante entender el mecanismo de derrame, ya sea biológico, social, ambiental o de otro tipo. Si no se sabe cuál es el mecanismo de derrame, no será posible elegir con precisión los grupos de comparación que son y no son afectados por los derrames. En segundo lugar, una evaluación con efectos de derrame requiere una recopilación de datos más amplia que una evaluación en la cual esa preocupación no existe: hay un grupo de comparación adicional (en el ejemplo anterior, los pueblos vecinos). Puede que también tengan que recopilarse datos sobre las otras unidades (en el ejemplo anterior, los adultos de los hogares objetivo para visitas relacionadas con la nutrición de los niños). En el recuadro 9.4 se analiza cómo los investigadores manejaron los efectos de derrame en una evaluación de un programa de transferencias condicionadas en México.

Recuadro 9.4: Evaluación de los efectos de derrame: transferencias condicionadas y derrames en México

Angelucci y De Giorgi (2009) analizaron los derrames en el programa Progresa, en México, que proporcionaba transferencias condicionadas a los hogares (véanse los recuadros 1.1 y 4.2). Los investigadores buscaban analizar si había riesgo compartido en los pueblos. Si los hogares compartían riesgo, los hogares elegibles podían transferir parte del efectivo a hogares no elegibles a través de préstamos o regalos.

El programa Progresa se implantó por fases a lo largo de dos años, y se seleccionaron aleatoriamente 320 pueblos para recibir las transferencias de efectivo en 1998, y 186 en 1999. Por lo tanto, entre 1998 y 1999 había 320 pueblos de tratamiento y 186 pueblos de comparación. En los pueblos de tratamiento, la elegibilidad de un hogar para las

Continúa en la página siguiente.

Recuadro 9.4: Evaluación de los efectos de derrame: transferencias condicionadas y derrames en México *(continúa)*

transferencias de Progresa estaba determinada por el nivel de pobreza y se contaba con los datos del censo de ambos grupos. Esto creó cuatro subgrupos dentro de la muestra: poblaciones elegibles y no elegibles en los pueblos de tratamiento y comparación. Suponiendo que el programa no afectaba indirectamente a los pueblos de comparación, los hogares no elegibles en dichos pueblos constituían un contrafactual válido para los hogares no elegibles en los pueblos de tratamiento, con el objetivo de estimar el efecto de derrame en los hogares no elegibles dentro de los pueblos de comparación.

Los autores encontraron evidencia de derrames positivos en el consumo. El consumo de alimentos de los adultos aumentó cerca de un 10% al mes en los hogares no elegibles de los pueblos de tratamiento. Esto equivalía a alrededor de la mitad del incremento promedio del consumo de alimentos de los hogares elegibles. Los resultados también apoyaron la hipótesis de riesgo compartido en los pueblos. Los hogares no elegibles en los pueblos de tratamiento recibieron más préstamos y transferencias de los amigos y la familia que los hogares no elegibles en los pueblos de comparación. Esto implica que el efecto de derrame funcionó a través de los mercados de seguro y de crédito locales.

A partir de estos resultados, Angelucci y De Giorgi llegaron a la conclusión de que las anteriores evaluaciones de Progresa subestimaban el impacto del programa en un 12% porque no tenían en cuenta los efectos indirectos en los hogares no elegibles en los pueblos de tratamiento.

Fuente: Angelucci y De Giorgi (2009).

El desgaste

El *sesgo del desgaste* es otro problema habitual que afecta a las evaluaciones, ya sea con el método de asignación aleatoria, de regresión discontinua o de diferencias en diferencias. El *desgaste* se produce cuando partes de la muestra "desaparecen" a lo largo del tiempo y los investigadores no pueden encontrar a todos los miembros iniciales de los grupos de tratamiento y comparación en las encuestas o en los datos de seguimiento. Por ejemplo, de los 2.500 hogares encuestados en la línea de base, los investigadores pueden encontrar solo 2.300 en una encuesta de seguimiento dos años después. Si intentan volver a realizar la encuesta al mismo grupo, por ejemplo, 10 años después, puede que encuentren incluso menos hogares originales.

El desgaste se puede producir por diferentes motivos. Por ejemplo, puede que los miembros de los hogares o incluso familias enteras se muden a otro pueblo, ciudad, región, o incluso país. En un ejemplo reciente, una encuesta de seguimiento realizada 22 años después en Jamaica indicó que el 18% de la

muestra había emigrado (véase el recuadro 9.5). En otros casos, los encuestados ya no estaban dispuestos a responder a una segunda encuesta. También ocurrió que los conflictos y la falta de seguridad en la zona impidieron que el equipo de investigación llevara a cabo una encuesta en algunas localidades incluidas en la línea de base.

Recuadro 9.5: El desgaste en estudios con seguimiento a largo plazo: desarrollo infantil temprano y migración en Jamaica

El desgaste puede ser especialmente problemático cuando han pasado muchos años entre las encuestas de línea de base y seguimiento. En 1986 un equipo de la University of West Indies inició un estudio para medir los resultados a largo plazo de una intervención en la primera infancia en Jamaica. En 2008, cuando los participantes originales tenían 22 años, se llevó a cabo un seguimiento. Fue difícil encontrar la pista de todos los participantes del estudio original.

La intervención consistió en un programa de dos años que ofreció estimulación psicosocial y suplementos nutricionales a niños pequeños con problemas de crecimiento en Kingston, Jamaica. Un total de 129 niños fueron asignados aleatoriamente a una de tres ramas de tratamiento o a un grupo de comparación. Los investigadores también encuestaron a 84 niños sin problemas de crecimiento para un segundo grupo de comparación. En el seguimiento, los investigadores pudieron realizar una segunda encuesta con casi el 80% de los participantes. No se recogió evidencia de desgaste selectivo en el conjunto de la muestra, lo que significa que no había diferencias significativas en las características de línea de base de aquellos que podían ser encuestados a los 22 años, comparados con aquellos que no podían ser encuestados. Sin embargo, al considerarse el subgrupo de niños que se habían convertido en trabajadores migrantes, se observaron señales de desgaste selectivo. De los 23 trabajadores migrantes, nueve habían abandonado la muestra y una parte significativamente mayor de aquellos pertenecía al grupo de tratamiento. Esto implicaba que el tratamiento estaba asociado con la migración. Dado que los trabajadores migrantes suelen ganar más que aquellos que permanecen en Jamaica, esto hizo difícil la formulación de estimaciones de impacto.

Para tratar los sesgos potenciales del desgaste entre los trabajadores migrantes, los investigadores utilizaron técnicas econométricas. Predijeron los ingresos de los trabajadores migrantes que habían abandonado la muestra mediante una regresión de mínimos cuadrados ordinarios utilizando como factores determinantes la condición de tratamiento, el sexo y la migración. Con estas predicciones en la estimación de impacto, llegaron a la conclusión de que los resultados del programa eran impresionantes. La intervención en la primera infancia había aumentado los ingresos del grupo de tratamiento en un 25%. Este efecto era lo bastante grande para que el grupo de tratamiento con problemas de crecimiento alcanzara al grupo de comparación sin problemas de crecimiento 20 años más tarde.

Fuentes: Gertler et al. (2014); Grantham-McGregor et al. (1991).

El desgaste puede ser problemático por dos motivos. En primer lugar, la muestra de seguimiento quizá ya no represente adecuadamente a la población de interés. Recuérdese que cuando se elige la muestra, en el momento de la asignación aleatoria, se hace de manera que represente de forma apropiada a la población de interés. En otras palabras, se escoge una muestra que tiene validez externa para la población de interés. Si la encuesta o la recopilación de datos de seguimiento se ve limitada por un desgaste considerable, debería ser preocupante que la muestra de seguimiento represente solo a un subconjunto específico de la población de interés. Por ejemplo, si las personas de mayor nivel educativo de la muestra original también son las que emigran, la encuesta de seguimiento ignoraría a aquellas personas con estudios y ya no representaría adecuadamente a la población de interés, que incluía a esas personas.

En segundo lugar, puede que la muestra de seguimiento ya no esté equilibrada entre el grupo de tratamiento y de comparación. Supóngase que se intenta evaluar un programa que quiere mejorar la educación de las niñas y que es más probable que las niñas con estudios se muden a la ciudad a buscar un empleo. Entonces, la encuesta de seguimiento podría mostrar un alto desgaste desproporcionado en el grupo de tratamiento, en relación con el grupo de comparación. Esto podría afectar la validez interna del programa, es decir, al contrastar las unidades de tratamiento y comparación que se encuentran en el seguimiento, ya no se podrá dar una estimación precisa del impacto del programa.

Si durante las encuestas de seguimiento se halla desgaste, los siguientes dos pasos pueden ayudar a evaluar el alcance del problema. Primero, verifíquese si las características de línea de base de las unidades que abandonaron la muestra son estadísticamente iguales a las características de línea de base de las unidades que fueron encuestadas con éxito la segunda vez. Siempre que las características de línea de base de ambos grupos no sean estadísticamente diferentes, la nueva muestra debería seguir representando a la población de interés.

Segundo, verifíquese si la tasa de desgaste del grupo de tratamiento es similar a la tasa de desgaste del grupo de comparación. Si ambas son significativamente diferentes, surge la preocupación de que la muestra ya no sea válida, y quizá deban utilizarse diversas técnicas estadísticas para intentar corregir esto. Un método habitual es la *ponderación por probabilidad inversa*, un método que *repondera* estadísticamente los datos (en este caso, los datos de seguimiento) para corregir el hecho de que una parte de los encuestados originales está ausente. El método formula una reponderación de la muestra de seguimiento de modo que tenga un aspecto similar a la muestra de línea de base.[1]

Programación en el tiempo y persistencia de los efectos

Los canales de transmisión entre insumos, actividades, productos y resultados pueden tener lugar de inmediato, pronto o después de un período de tiempo, y suelen estar estrechamente relacionados con los cambios en el comportamiento humano. En el capítulo 2 se ponía de relieve la importancia de pensar en estos canales y planificar correspondientemente antes de que comenzara la intervención, así como de desarrollar una cadena causal clara para el programa que se esté evaluando. En aras de la sencillez, nos hemos abstraído de los problemas relacionados con la programación en el tiempo. Sin embargo, es fundamental considerar estos aspectos cuando se diseña una evaluación.

En primer lugar, los programas no necesariamente se vuelven plenamente efectivos justo después de su inicio (King y Behrman, 2009). Los administradores de un programa necesitan tiempo para que éste comience a funcionar, y puede ser que los beneficiarios no vean los frutos de inmediato porque los cambios de conducta requieren tiempo, y puede ser que las instituciones tampoco modifiquen su comportamiento con rapidez. Por otro lado, una vez que las instituciones y los beneficiarios cambian ciertas conductas, puede ocurrir que estas se mantengan aun cuando se suspenda el programa. Por ejemplo, un programa que incentiva a los hogares a separar y reciclar la basura y ahorrar energía puede seguir siendo efectivo después de que se eliminen los incentivos, si consigue cambiar las normas de los hogares en el manejo de la basura y la energía. Cuando se diseña una evaluación, hay que tener mucho cuidado (y ser realistas) para definir cuánto podría tardar el programa en alcanzar su plena efectividad. Puede que sea necesario llevar a cabo diversas encuestas de seguimiento para medir el impacto del programa a lo largo del tiempo, o incluso después de que el programa se interrumpa. El recuadro 9.6 presenta el caso de una evaluación donde algunos efectos solo se hicieron visibles después de suspendida la intervención inicial.

Recuadro 9.6: Evaluación de los efectos a largo plazo: subsidios y adopción de redes antimosquitos tratadas con insecticidas en Kenia

Dupas (2014) diseñó una evaluación de impacto para medir los impactos tanto de corto como de largo plazo de diferentes esquemas de subsidios en la demanda de redes antimosquitos tratadas con insecticidas (ITN, por sus siglas en inglés, *insecticide treated bed nets*) en Busia, Kenia. Utilizando un experimento de dos fases donde intervenía la fijación de precios, Dupas asignó aleatoriamente hogares a diversos niveles de subsidios para un nuevo tipo de ITN.

Continúa en la página siguiente.

Recuadro 9.6: Evaluación de los efectos a largo plazo: subsidios y adopción de redes antimosquitos tratadas con insecticidas en Kenia *(continúa)*

Un año después, todos los hogares en un subconjunto de pueblos tuvieron la oportunidad de comprar la misma red. Esto permitió a los investigadores medir la disponibilidad de los hogares a pagar por las ITN y cómo esta disponibilidad cambiaba en función del subsidio recibido en la primera fase del programa.

En general, los resultados indicaron que un subsidio único tenía impactos significativamente positivos en la adopción de ITN y la disponibilidad para pagar a largo plazo. En la primera fase del experimento, Dupas observó que los hogares que recibían un subsidio que reducía el precio de la ITN de US$3,80 a US$0,75 tenían un 60% más de probabilidades de comprarla. Cuando la ITN se ofreció gratis, la tasa de adopción aumentó al 98%. A largo plazo, las tasas de adopción más altas se tradujeron en una mayor disponibilidad a pagar, dado que los hogares vieron los beneficios de tener una ITN. Aquellos que recibieron uno de los subsidios más grandes en la primera fase tenían tres veces más probabilidades de comprar otra ITN en la segunda fase a más del doble del precio.

Los resultados de este estudio implican que se produce un efecto de aprendizaje en las intervenciones en ITN. Esto señala que es importante considerar los impactos de las intervenciones a largo plazo, así como dar a conocer la persistencia de los efectos.

Fuente: Dupas (2014).

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos a más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).

Nota

1. Un método estadístico más avanzado sería estimar "límites nítidos" en los efectos del tratamiento (véase Lee, 2009).

Referencias bibliográficas

Angelucci, M. y G. De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption." *American Economic Review* 99 (1): 486–508.

Angelucci, M. y V. Di Maro. 2015. "Programme Evaluation and Spillover Effects." *Journal of Development Effectiveness* (doi: 10.1080/19439342.2015.1033441).

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot y P. Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128 (2): 531–80.

Dupas, P. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82 (1): 197–228.

Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang y S. Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Grantham-McGregor, S., C. Powell, S. Walker y J. Himes. 1991. "Nutritional Supplementation, Psychosocial Stimulation and Development of Stunted Children: The Jamaican Study." *Lancet* 338: 1–5.

King, E. M. y J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 (1): 55–82.

Kremer, M. y E. Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

Landsberger, H. A. 1958. *Hawthorne Revisited.* Ithaca, NY: Cornell University Press.

Lee, D. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3):1071–1102.

Levitt, S. D. y J. A. List. 2009. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." Documento de trabajo NBER 15016. Cambridge, MA: National Bureau of Economic Research.

Saretsky, G. 1972. "The OEO P.C. Experiment and the John Henry Effect." *Phi Delta Kappan* 53: 579–81.

CAPÍTULO 10

Evaluación de programas multifacéticos

Evaluación de programas que combinan diversas opciones de tratamiento

Hasta ahora, se han analizado programas con un solo tipo de tratamiento. Sin embargo, muchas cuestiones relevantes relacionadas con las políticas se plantean en programas multifacéticos, es decir, que combinan varias opciones de tratamiento.[1] Los responsables de formular políticas pueden estar interesados en saber no solo si el programa funciona o no, sino también si funciona mejor o tiene un costo menor que otro programa. Por ejemplo, si se quiere aumentar la asistencia a la escuela, ¿es más eficaz orientar las intervenciones a la demanda (como las transferencias condicionadas a las familias) o a la oferta (como mayores incentivos para los profesores)? Y si se introducen las dos intervenciones conjuntamente, ¿funcionan mejor que cada una por su cuenta?, ¿son complementarias? Si la costo-efectividad es una prioridad, puede preguntarse perfectamente cuál es el nivel óptimo de los servicios que debe prestar el programa. Por ejemplo, ¿cuál es la duración óptima de un programa de capacitación para el empleo? ¿Un programa de seis meses contribuye más que un programa de tres meses a que los participantes encuentren empleo? De ser así, ¿la diferencia es lo suficientemente grande para justificar los recursos adicionales necesarios para un programa de seis meses? Por último, a los responsables de políticas les puede interesar cómo alterar un

programa existente para hacerlo más efectivo, y quizá quieran probar diversos mecanismos con el fin de encontrar cuál(es) funciona(n) mejor.

Además de estimar el impacto de una intervención sobre un resultado de interés, las evaluaciones de impacto pueden ayudar a responder preguntas más generales, como las siguientes:

- ¿Cuál es el impacto de un tratamiento en comparación con otro? Por ejemplo, ¿cuál es el impacto en el desarrollo cognitivo de los niños de un programa que ofrece capacitación a los padres, en comparación con una intervención sobre nutrición?
- ¿El impacto conjunto de un primer y un segundo tratamiento es mayor que la suma de los dos impactos? Por ejemplo, ¿el impacto de la intervención de capacitación de padres y la intervención sobre nutrición es mayor, menor o igual que la suma de los efectos de cada una de las intervenciones?
- ¿Cuál es el impacto de un tratamiento de alta intensidad en comparación con un tratamiento de menor intensidad? Por ejemplo, ¿cuál es el efecto en el desarrollo cognitivo de niños con retraso en el crecimiento si un trabajador social los visita en su casa cada dos semanas, en lugar de visitarlos una vez al mes?

Este capítulo ofrece ejemplos de diseños de evaluaciones de impacto para dos tipos de programas multifacéticos: los que tienen múltiples niveles del mismo tratamiento y los que tienen múltiples tratamientos. Primero se analiza cómo diseñar una evaluación de impacto de un programa con varios niveles de tratamiento. Después, se examinan los diferentes tipos de impactos de un programa con múltiples tratamientos. Para este análisis se supone que se usará un método de asignación aleatoria, aunque puede generalizarse a otros métodos.

Evaluación de programas con diferentes niveles de tratamiento

Diseñar una evaluación de impacto para un programa con niveles variables de tratamiento es relativamente fácil. Imagínese que se intenta evaluar el impacto de un programa con dos niveles de tratamiento: alto (por ejemplo, visitas cada dos semanas) y bajo (visitas mensuales). Se quiere evaluar el impacto de ambas opciones, y saber cuánto afectan a los resultados esas visitas adicionales. Para ello, se puede organizar un sorteo de modo de decidir quién recibe el nivel alto de tratamiento, quién recibe el nivel bajo de tratamiento y a quién se asigna al grupo de comparación (el gráfico 10.1 ilustra este proceso).

Como es habitual en la asignación aleatoria, el primer paso consiste en definir la población de unidades elegibles para el programa. El segundo, en seleccionar una muestra aleatoria de unidades que se incluirá en la evaluación,

Gráfico 10.1 Pasos para la asignación aleatoria de dos niveles de tratamiento

la denominada muestra de evaluación. Una vez que se cuente con la muestra de evaluación, en el tercer paso se asignarán aleatoriamente unidades al grupo que recibe un nivel alto de tratamiento, al grupo que recibe el nivel bajo de tratamiento o al grupo de comparación. Como resultado de la asignación aleatoria a múltiples niveles de tratamiento, se habrán creado tres grupos distintos:

- El grupo A es el grupo de comparación.
- El grupo B recibe el nivel bajo de tratamiento.
- El grupo C recibe el nivel alto de tratamiento.

Cuando se implementa correctamente, la asignación aleatoria garantiza que los tres grupos sean similares. Por lo tanto, se puede estimar el impacto del nivel alto de tratamiento mediante la comparación del resultado promedio del grupo C con el resultado promedio del grupo A. También se puede estimar el nivel bajo de tratamiento comparando el resultado promedio del grupo B con el del grupo A. Finalmente, se puede evaluar si el nivel alto de tratamiento tiene un mayor impacto que el nivel bajo de tratamiento comparando los resultados promedio de los grupos B y C.

La estimación del impacto de un programa con más de dos niveles de tratamiento seguirá la misma lógica. Si existen tres niveles de tratamiento, el proceso de asignación aleatoria creará tres grupos de tratamiento diferentes, además de un grupo de comparación. En general, con *n* niveles de tratamiento, habrá *n* grupos de tratamiento, más un grupo de comparación. En los recuadros 10.1 y 10.2 se presentan ejemplos de evaluaciones de impacto que prueban modalidades de diferentes intensidades u opciones de tratamientos múltiples.

Concepto clave

Al evaluar programas con *n* diferentes niveles de tratamiento, debe haber *n* grupos de tratamiento más un grupo de comparación.

Recuadro 10.1: Prueba de la intensidad de un programa para mejorar la adhesión a un tratamiento antirretroviral

Pop-Eleches et al. (2011) utilizaron un diseño multinivel transversal para evaluar el impacto del uso de mensajes SMS como recordatorios para la adhesión de los pacientes con VIH/SIDA a la terapia antirretroviral en una clínica rural de Kenia. El estudio varió la intensidad del tratamiento en dos dimensiones: la frecuencia con que se enviaban los mensajes a los pacientes (a diario o semanalmente) y la extensión de los mensajes (breves o largos). Los mensajes breves tenían solo un recordatorio ("Este es un recordatorio para usted"), mientras que los mensajes largos incluían un recordatorio y una frase de aliento ("Este es un recordatorio. Sea fuerte y valiente, nos preocupamos por usted"). Se asignó un total de 531 pacientes a uno de los cuatro grupos de tratamiento o al grupo de comparación. Los grupos de tratamiento consistían en: mensajes semanales breves, mensajes semanales largos, mensajes diarios breves o mensajes diarios largos.

Una tercera parte de la muestra se asignó al grupo de control y las otras dos terceras partes se asignaron por igual a cada uno de los cuatro grupos de intervención. Se generó una secuencia de números aleatorios entre 0 y 1. Cuatro intervalos iguales entre 0 y 2/3 correspondían a los cuatro grupos de intervención, mientras que el intervalo de valor de 2/3 a 1 correspondía al grupo de control.

Los investigadores concluyeron que los mensajes semanales aumentaban el porcentaje de pacientes con un 90% de adhesión a la terapia antirretroviral en alrededor de un 13%-16%, en comparación con la ausencia de mensajes. Estos mensajes semanales también eran efectivos para reducir la frecuencia de las interrupciones del tratamiento, que –según se ha demostrado– constituyen una causa importante del fracaso por resistencia al tratamiento en contextos de recursos limitados. Contrariamente a las expectativas, añadir palabras de aliento en los mensajes más largos no era más efectivo que un mensaje breve o ningún mensaje.

Los investigadores también descubrieron que si bien los mensajes semanales mejoraban la adhesión, los mensajes diarios no lo hacían, pero no fueron capaces de distinguir por qué los primeros eran los más efectivos. Es posible que esta conclusión se explique gracias a la habituación, o la menor respuesta ante un estímulo repetido con frecuencia, o puede que los pacientes sencillamente opinaran que los mensajes diarios eran intrusivos.

Cuadro B10.1.1 Resumen del diseño del programa

Grupo	Tipo de mensaje	Frecuencia del mensaje	Nº de pacientes
1	Solo recordatorio	Semanal	73
2	Recordatorio + aliento	Semanal	74
3	Solo recordatorio	Diario	70
4	Recordatorio + aliento	Diario	72
5	Ninguno (grupo de comparación)	Ninguna	139

Fuente: Pop-Eleches et al. (2011).

Recuadro 10.2: Pruebas de alternativas de los programas para monitorear la corrupción en Indonesia

En Indonesia, Olken (2007) utilizó un diseño transversal para probar diferentes métodos con el fin de controlar la corrupción, desde una estrategia de vigilancia de arriba hacia abajo hasta una supervisión comunitaria más de base. El autor recurrió a una metodología de asignación aleatoria en más de 600 comunidades que estaban construyendo carreteras como parte de un proyecto nacional de mejora de infraestructura.

Uno de los tratamientos múltiples consistió en seleccionar de manera aleatoria algunas comunidades para informarles que su proyecto de construcción sería auditado por un funcionario público. Luego, para poner a prueba la participación comunitaria en la supervisión, los investigadores implementaron dos intervenciones. Distribuyeron invitaciones a reuniones comunitarias para la rendición de cuentas y repartieron formularios para presentar comentarios de manera anónima. Para medir los niveles de corrupción, un equipo independiente de ingenieros y topógrafos tomó muestras básicas de las nuevas carreteras, estimó el costo de los materiales usados y comparó sus cálculos con los presupuestos presentados.

Olken observó que el incremento de las auditorías públicas (desde una probabilidad de resultar auditado de alrededor del 4% hasta una probabilidad del 100%) redujo la pérdida de gastos en unos 8 puntos porcentuales (a partir de un 24%). El aumento de la participación de la comunidad en la supervisión tuvo un impacto sobre la pérdida de mano de obra pero no sobre la pérdida de gastos. Los formularios para comentarios solo resultaron eficaces cuando se distribuyeron entre los niños en la escuela para que se los entregaran a sus familias, y no cuando fueron entregados a los líderes comunitarios.

Fuente: Olken (2007).

Evaluación de múltiples intervenciones

Además de comparar varios niveles de tratamiento, también se pueden comparar opciones de tratamiento totalmente diferentes. De hecho, los responsables de las políticas prefieren comparar los méritos relativos de diferentes intervenciones, más que conocer solo el impacto de una intervención.

Imagínese que se propone evaluar el impacto en la matriculación escolar de un programa con dos intervenciones, transferencias condicionadas a las familias de los estudiantes y transporte gratuito en autobús a la escuela. Primero, es necesario conocer el impacto de cada intervención por separado. Este caso es prácticamente idéntico a aquel en que se prueban diferentes niveles de tratamiento de una intervención, a saber, en lugar de asignar aleatoriamente las unidades a niveles altos y bajos de tratamiento y al grupo de comparación, se les puede asignar de forma aleatoria a un grupo de

transferencias condicionadas, a un grupo de transporte gratuito en autobús y al grupo de comparación. En general, con *n* niveles de tratamiento, habrá *n* grupos de tratamiento, más un grupo de comparación.

Aparte de querer conocer el impacto de cada intervención por separado, puede que también se desee conocer si la combinación de los dos es mejor que la simple suma de los efectos individuales. Desde el punto de vista de los participantes, el programa está disponible en tres formas diferentes: solo transferencias condicionadas, únicamente transporte gratuito en autobús o una combinación de transferencias y transporte gratuito.

La asignación aleatoria para un programa con dos intervenciones es muy similar al proceso de un programa con una sola intervención. La principal diferencia es la necesidad de organizar varios sorteos independientes, en lugar de uno. Esto produce un *diseño cruzado*, a veces llamado diseño transversal. En el gráfico 10.2 se ilustra este proceso. Como en el caso anterior, en el primer paso se define la población de unidades elegibles para el programa. El segundo paso consiste en seleccionar una muestra aleatoria de unidades elegibles para formar la muestra de evaluación. Una vez obtenida la muestra de evaluación, en el tercer paso se asignan aleatoriamente sus unidades a un grupo de tratamiento y a un grupo de control. En el cuarto paso, se lleva a cabo un segundo sorteo para asignar de forma aleatoria una subserie del grupo de tratamiento a fin de que reciba la segunda intervención. Por último, en el quinto paso se realiza otro sorteo para asignar una subserie del grupo de comparación inicial a fin de que reciba la segunda intervención, mientras que la otra subserie se mantiene como un conjunto puro de comparación.[2]

Gráfico 10.2 Pasos para la asignación aleatoria de dos intervenciones

Como consecuencia de la asignación aleatoria a los dos tratamientos, se habrán creado cuatro grupos, como se muestra en el gráfico 10.3.

- El grupo A recibe ambas intervenciones (transferencias condicionadas y transporte en autobús).
- El grupo B recibe la primera intervención pero no la segunda (solo transferencias condicionadas).
- El grupo C no recibe la primera intervención pero sí la segunda (solo el transporte en autobús).
- El grupo D no recibe ni la primera ni la segunda intervención, y constituye el grupo de comparación puro.

Cuando se implementa correctamente, la asignación aleatoria garantiza que los cuatro grupos sean similares. Por lo tanto, se puede estimar el impacto de la primera intervención comparando el resultado del grupo B (por ejemplo, la tasa de asistencia escolar) con el resultado del grupo puro de comparación, el grupo D. También se puede estimar el impacto de la segunda intervención comparando el resultado del grupo C con el resultado del grupo de comparación puro, el grupo D. Además, este diseño también permite comparar el impacto progresivo de recibir la segunda intervención cuando una unidad ya ha recibido la primera. La comparación de los resultados del

Gráfico 10.3 Diseño híbrido para un programa con dos intervenciones

grupo A y del grupo B determinará el impacto de la segunda intervención para aquellas unidades que ya han recibido la primera intervención. La comparación de los resultados de los grupos A y C determinará el impacto de la primera intervención en las unidades que ya han recibido la segunda intervención.

En la descripción anterior se ha usado el ejemplo de la asignación aleatoria para explicar la manera de diseñar una evaluación de impacto para un programa con dos intervenciones diferentes. Cuando un programa cuenta con más de dos intervenciones, se puede aumentar el número de sorteos y continuar subdividiendo la evaluación para formar grupos que reciben las diversas combinaciones de intervenciones. También se pueden implementar múltiples tratamientos y múltiples niveles de tratamiento. Aunque se amplíe el número de grupos, la teoría fundamental del diseño sigue siendo la misma que la descripta anteriormente.

Sin embargo, la evaluación de más de una o dos intervenciones generará dificultades prácticas tanto en la evaluación como en el funcionamiento del programa, ya que la complejidad del diseño incrementará exponencialmente el número de ramas de tratamiento. Para evaluar el impacto de una sola intervención se necesitan únicamente dos grupos, uno de tratamiento y otro de comparación. Para evaluar el impacto de dos intervenciones se necesitan cuatro grupos, tres de tratamiento y uno de comparación. Si se quisiera evaluar el impacto de tres intervenciones, incluidas todas las combinaciones posibles entre ellas, se necesitaría $2 \times 2 \times 2 = 8$ grupos en la evaluación. En general, en el caso de una evaluación que vaya a incluir todas las combinaciones posibles entre n intervenciones, se necesitarán 2^n grupos. Además, para poder distinguir los resultados de los grupos, cada grupo requiere un número suficiente de unidades de observación de modo de garantizar una potencia estadística suficiente. En la práctica, la detección de diferencias entre las ramas de la intervención puede exigir muestras más grandes que la comparación entre un grupo de tratamiento y un grupo de comparación puro. Si las dos ramas de tratamiento logran provocar cambios en los resultados deseados, se requerirán muestras más grandes para detectar las posibles diferencias menores entre los dos grupos.[3]

Concepto clave

Para que una evaluación estime el impacto de todas las posibles combinaciones entre n intervenciones diferentes, se requerirá un total de 2^n grupos de tratamiento y de comparación.

Por último, los diseños cruzados también se pueden utilizar en diseños de evaluación que combinan diversos métodos de evaluación. Las reglas operativas que rigen la asignación de cada tratamiento determinarán qué combinación de métodos debe usarse. Por ejemplo, puede ocurrir que el primer tratamiento se asigne sobre la base de una puntuación de elegibilidad, pero el segundo se asignará de manera aleatoria. En este caso, el diseño puede recurrir a un diseño de regresión discontinua para la primera intervención y a un método de asignación aleatoria para la segunda intervención.

Otros recursos

- Para material de apoyo relacionado con este libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para más información sobre el diseño de evaluaciones de impacto con múltiples opciones de tratamiento, véase A. Banerjee y E. Duflo (2009), "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Notas

1. Véase Banerjee y Duflo (2009), para un análisis más detallado.
2. Nótese que, en la práctica, es posible combinar los tres sorteos separados en uno solo y alcanzar el mismo resultado.
3. Probar el impacto de múltiples intervenciones también tiene una implicación más sutil: a medida que se incrementa el número de intervenciones o niveles de tratamiento que se contrastan unos con otros, se aumenta la probabilidad de encontrar un impacto en al menos una de las pruebas, aunque no haya impacto. En otras palabras, hay más probabilidades de encontrar un falso positivo. Para evitar esto, se deben ajustar las pruebas estadísticas de modo de dar cuenta de las pruebas de hipótesis múltiples. Los falsos positivos también se denominan errores de tipo II. Véase el capítulo 15 para más información sobre los errores de tipo II y referencias sobre las pruebas de hipótesis múltiples.

Referencias bibliográficas

Banerjee, A. y E. Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Olken, B. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–249.

Pop-Eleches, C., H. Thirumurthy, J. Habyarimana, J. Zivin, M. Goldstein, D. de Walque, L. MacKeen, J. Haberer, S. Kimaiyo, J. Sidle, D. Ngare y D. Bangsberg. 2011. "Mobile Phone Technologies Improve Adherence to Antiretroviral Treatment in a Resource-Limited Setting: A Randomized Controlled Trial of Text Message Reminders." *AIDS* 25 (6): 825–34.

Tercera parte

CÓMO IMPLEMENTAR UNA EVALUACIÓN DE IMPACTO

La tercera parte de este libro se centra en cómo implementar una evaluación de impacto: cómo seleccionar un método de evaluación de impacto compatible con las reglas operativas de un programa; cómo manejar una evaluación de impacto, lo cual incluye asegurar una sólida asociación entre los equipos de investigación y los responsables de las políticas, y gestionar el tiempo y el presupuesto de una evaluación; cómo garantizar que una evaluación sea a la vez ética y creíble, siguiendo los principios para trabajar con sujetos humanos y ciencia abierta; y cómo utilizar la evaluación de impacto para fundamentar las políticas públicas.

En el capítulo 11 se describe cómo usar las reglas operativas del programa como base para elegir un método de evaluación de impacto, a saber: los recursos

de que dispone un programa, el criterio para seleccionar a los beneficiarios y el calendario de la implementación. Se establece un marco de trabajo sencillo para determinar cuál de las metodologías de evaluación de impacto presentadas en la segunda parte es la más adecuada para un determinado programa, de acuerdo con sus reglas operativas. El capítulo también trata de cómo el mejor método es el que requiere los supuestos más débiles y tiene la menor cantidad de requisitos de datos en el contexto de las reglas operativas.

En el capítulo 12 se analiza la relación entre los equipos de investigación y de políticas públicas y sus respectivos roles. También se aborda la diferencia entre independencia y ausencia de sesgo, y se destacan ámbitos que pueden revelarse como sensibles en la realización de una evaluación de impacto. El capítulo ofrece orientación sobre cómo gestionar las expectativas de las partes interesadas y destaca algunos de los riesgos habituales presentes en las evaluaciones de impacto, así como sugerencias sobre cómo gestionar esos riesgos. Concluye con una visión general de cómo gestionar las actividades de evaluación de impacto, entre ellas la creación del equipo de evaluación, el calendario de la evaluación, el presupuesto y la recaudación de fondos.

El capítulo 13 proporciona una visión general de la ética y la ciencia de la evaluación de impacto, lo que incluye la importancia de no negar los beneficios a los beneficiarios elegibles en aras de la evaluación; cómo aplicar principios básicos de investigación ética con sujetos humanos; el rol de las juntas de revisión institucional que aprueban y monitorean la investigación con dichos sujetos; y la importancia de practicar la ciencia abierta, lo cual implica registrar las evaluaciones y divulgar públicamente los datos para otras investigaciones y para replicar los resultados.

El capítulo 14 presenta visiones novedosas sobre cómo utilizar las evaluaciones de impacto para fundamentar las políticas públicas, lo que abarca consejos sobre cómo destacar la relevancia de los resultados, un debate sobre el tipo de productos que las evaluaciones de impacto pueden y deben elaborar, y orientación sobre cómo producir y divulgar los hallazgos para maximizar el impacto de las políticas.

CAPÍTULO 11

Elección de un método de evaluación de impacto

¿Qué método usar en un determinado programa?

La clave para identificar el impacto causal de un programa consiste en encontrar un grupo de comparación válido para estimar el contrafactual y responder a la pregunta de interés de la política pública. En la segunda parte de este volumen se abordaron diversos métodos, entre ellos la asignación aleatoria, las variables instrumentales, la regresión discontinua, las diferencias en diferencias y el pareamiento. En este capítulo, se analizará la pregunta relacionada con qué método elegir para un determinado programa que se quiera evaluar.

En primer lugar, se muestra que las reglas operativas del programa constituyen una clara orientación para encontrar grupos de comparación y, por lo tanto, para hallar el método más adecuado para su contexto de políticas. Un principio general es que si las reglas operativas de un programa están bien definidas, pueden ayudar a determinar cuál es el método más adecuado para evaluar ese programa concreto.

En segundo lugar, los métodos introducidos en la segunda parte tienen diferentes requisitos de datos y se basan en diferentes supuestos fundamentales. Algunos métodos requieren supuestos más fuertes que otros para estimar con precisión los cambios en los resultados

"causados" por la intervención. En general, se prefiere el método que requiere los supuestos más débiles y tiene la menor cantidad de requisitos de datos en el contexto de las reglas operativas.

Por último, se examina cómo elegir la unidad de intervención. Por ejemplo, ¿el programa se asignará a nivel individual, o a un nivel superior, como las comunidades o los distritos? En general, conviene elegir la unidad de intervención factible más pequeña dentro de las limitaciones operativas.

Cómo las reglas operativas de un programa pueden contribuir a elegir un método de evaluación de impacto

Concepto clave

Las reglas operativas de un programa determinan qué método de evaluación de impacto es el más adecuado para evaluar el programa, no a la inversa.

Uno de los principales mensajes de este libro es que se pueden usar las reglas operativas de un programa para encontrar grupos de comparación válidos, en la medida en que las reglas operativas del programa estén bien definidas. De hecho, dichas reglas brindan orientación en cuanto al método más adecuado para evaluar ese programa concreto. Las reglas operativas del programa son las que pueden y deben regir en el método de evaluación, no a la inversa. La evaluación no debería cambiar drásticamente elementos clave de las reglas de asignación del programa bien definidas en aras de un diseño de evaluación más claro.

Las reglas operativas más pertinentes para el diseño de la evaluación son aquellas que identifican quiénes son elegibles para el programa y cómo se seleccionan para que participen. Los grupos de comparación provienen de aquellos sujetos elegibles pero que no pueden incorporarse al programa en un determinado momento (por ejemplo, cuando los recursos son limitados y hay exceso de demanda), o de aquellos que se encuentran cerca de un umbral de elegibilidad para participar en el programa.

Concepto clave

Cuando se diseñan evaluaciones de impacto, casi siempre se pueden encontrar grupos de comparación válidos si las reglas operativas para seleccionar a los beneficiarios son equitativas, transparentes y están sujetas a rendición de cuentas.

Principios de las reglas de asignación al programa bien definidas

Al diseñar las evaluaciones de impacto, siempre se pueden encontrar grupos de comparación válidos si las reglas operativas para seleccionar a los beneficiarios son equitativas, transparentes y están sujetas a rendición de cuentas:

- Las reglas *equitativas* de asignación al programa clasifican o priorizan la elegibilidad en función de un indicador de las necesidades acordado comúnmente, o estipulan que a todos se les ofrezcan los beneficios del programa, o que al menos tengan iguales posibilidades de que les ofrezcan los beneficios.

- Las reglas de asignación al programa se divulgan y son *transparentes*, de modo que las partes externas las acepten implícitamente y puedan comprobar que en efecto hay un seguimiento. Las reglas transparentes deberían ser cuantificables y fácilmente observables.
- Las reglas sujetas a *rendición de cuentas* son responsabilidad de los funcionarios del programa y su implementación es la base del desempeño en el trabajo o de las recompensas de esos funcionarios.

Las reglas operativas de elegibilidad son transparentes y están sujetas a rendición de cuentas cuando los programas utilizan criterios cuantificables que pueden ser verificados por organizaciones externas y cuando hacen públicos dichos criterios. La equidad, la transparencia y la rendición de cuentas aseguran que los criterios de elegibilidad sean verificables cuantitativamente y estén realmente implementados según su diseño. Como tales, estos principios de buena gobernanza mejoran la probabilidad de que el programa realmente beneficie a la población focalizada y también constituyen la clave de una evaluación exitosa. Si las reglas no son cuantificables y verificables, el equipo de evaluación tendrá dificultades para asegurar que la asignación a los grupos de tratamiento y comparación se produzca siguiendo el diseño o, como mínimo, documentando cómo sucedió en la práctica. Si los miembros del equipo de evaluación no pueden verificar la asignación en la práctica, no pueden analizar correctamente los datos para calcular los impactos. Entender las reglas de asignación del programa es fundamental para seleccionar un método de evaluación adecuado.

Cuando las reglas operativas incumplen cualquiera de estos tres principios de buena gobernanza, surgen dificultades tanto para crear un programa bien diseñado como para llevar a cabo la evaluación. Es difícil encontrar grupos de comparación válidos si las reglas que determinan la elegibilidad y la selección de los beneficiarios no son equitativas ni transparentes, ni están sujetas a rendición de cuentas. En este caso, el diseño de una evaluación de impacto puede requerir aclaraciones y ajustes en el funcionamiento del programa. Sin embargo, si las reglas están bien definidas, el método de evaluación de impacto se puede elegir sobre la base de las reglas existentes de asignación del programa, como se explica a continuación con más detalle.

Reglas operativas clave

Las reglas operativas suelen definir cuáles son los beneficios del programa, cómo se financian y se distribuyen estos beneficios y de qué modo el programa selecciona a los beneficiarios. Las reglas que gobiernan

los programas y la selección de los beneficiarios son clave para encontrar grupos de comparación válidos. Las reglas que gobiernan la selección de los beneficiarios comprenden la elegibilidad, las reglas de asignación en el caso de recursos limitados y el orden de incorporación de los beneficiarios a lo largo del tiempo. Más específicamente, las reglas clave que generan una hoja de ruta para encontrar grupos de comparación corresponden a tres preguntas operativas fundamentales en relación con los recursos de los que dispone un programa, los criterios de elegibilidad y el calendario de la implementación:

1. *Recursos disponibles*: ¿El programa tiene suficientes recursos para implementarse a escala y atender a todos los beneficiarios elegibles? Los gobiernos y las organizaciones no gubernamentales (ONG) no siempre tienen suficientes recursos para proporcionar los servicios del programa a todos los que son elegibles y se postulan a los programas. En ese caso, el gobierno u ONG debe decidir cuáles son los postulantes elegibles que recibirán los beneficios del programa y cuáles quedarán excluidos. En muchas ocasiones, los programas se limitan a regiones geográficas específicas o a un número restringido de comunidades, aun cuando pueda haber beneficiarios elegibles en otras regiones o comunidades.

2. *Criterios de elegibilidad*: ¿Quién es elegible para recibir los beneficios del programa? ¿La asignación del programa se basa en un umbral de elegibilidad, o está disponible para todos? Las escuelas públicas y la atención primaria de salud suelen ser de carácter universal. Muchos programas utilizan reglas operativas de elegibilidad que dependen de una clasificación continua y un umbral definido. Por ejemplo, los sistemas de jubilación establecen una edad límite por encima de la cual las personas de edad avanzada son elegibles. Los programas de transferencias condicionadas suelen clasificar a los hogares a partir de su nivel estimado de pobreza y los hogares por debajo de un umbral de pobreza predeterminado se consideran elegibles.

3. *Calendario de implementación*: ¿Los beneficiarios potenciales se inscriben en el programa todos al mismo tiempo o por fases? A menudo, las limitaciones administrativas y de recursos impiden que los gobiernos y las ONG proporcionen beneficios de forma inmediata a toda la población elegible. Deben implementar sus programas a lo largo del tiempo y, por lo tanto, decidir quién es objeto de los beneficios primero y quién se incorpora más tarde. Un enfoque común consiste en ampliar un programa en fases geográficamente, a lo largo del tiempo, incorporando a todos los beneficiarios elegibles en una comunidad o región antes de pasar a la siguiente.

Creación de grupos de comparación a partir de las reglas operativas

Al diseñar evaluaciones de impacto prospectivas, la respuesta a las tres preguntas operativas determina en gran parte el método de evaluación de impacto más adecuado para un cierto programa. El cuadro 11.1 recoge los

Cuadro 11.1 Relación entre las reglas operativas de un programa y los métodos de evaluación de impacto

		Exceso de demanda del programa (recursos limitados)		No exceso de demanda del programa (recursos suficientes)	
	Criterios de elegibilidad	(1) Índice continuo y umbral de elegibilidad	(2) Sin índice continuo y umbral de elegibilidad	(3) Índice continuo y umbral de elegibilidad	(4) Sin índice continuo y umbral de elegibilidad
Calendario de implementación	(A) Implementación en fases	Celda A1 Asignación aleatoria (capítulo 4) DRD (capítulo 6)	Celda A2 Asignación aleatoria (capítulo 4) Variables instrumentales (promoción aleatoria) (capítulo 5) DD (capítulo 7) DD con pareamiento (capítulo 8)	Celda A3 Asignación aleatoria en fases (capítulo 4) DRD (capítulo 6)	Celda A4 Asignación aleatoria en fases (capítulo 4) Variables instrumentales (promoción aleatoria para participación temprana) (capítulo 5) DD (capítulo 7) DD con pareamiento (capítulo 8)
	(B) Implementación inmediata	Celda B1 Asignación aleatoria (capítulo 4) DRD (capítulo 6)	Celda B2 Asignación aleatoria (capítulo 4) Variables instrumentales (promoción aleatoria) (capítulo 5) DD (capítulo 7) DD con pareamiento (capítulo 8)	Celda B3 DRD (capítulo 6)	Celda B4 Si participación no es plena: Variables instrumentales (promoción aleatoria) (capítulo 5) DD (capítulo 7) DD con pareamiento (capítulo8)

Nota: DD = diferencias en diferencias; DRD = diseño de regresión discontinua.

posibles grupos de comparación en relación con las reglas operativas específicas del programa y las tres preguntas operativas fundamentales relacionadas con los recursos disponibles, las reglas de elegibilidad y el calendario de implementación. Las columnas se dividen en función de si el programa tiene o no recursos suficientes para eventualmente cubrir a todos los beneficiarios elegibles potenciales (*recursos disponibles*) y, además, se subdividen en programas que tienen un *ranking* y un límite de elegibilidad continuos y aquellos que no los tienen (*criterios de elegibilidad*). Las filas se dividen en implementación en fases o implementación inmediata del programa (*calendario de implementación*). En cada celda se recogen las fuentes potenciales de grupos de comparación válidos, junto con el capítulo relacionado en que se trata en la segunda parte. Cada celda tiene un rótulo con un índice: la letra inicial señala la columna en el cuadro (A, B) y el número que sigue indica la columna (1-4). Por ejemplo, la celda A1 se refiere a la celda de la primera fila y la primera columna del cuadro. Así, la celda A1 identifica los métodos de evaluación más adecuados para los programas que tienen recursos limitados, que tienen criterios de elegibilidad y se desarrollan en fases.

La mayoría de los programas debe implementarse en fases a lo largo del tiempo debido ya sea a limitaciones financieras o a problemas logísticos y administrativos. Este grupo o categoría cubre la primera fila del cuadro (celdas A1, A2, A3 y A4). En este caso, la regla operativa equitativa, transparente y sujeta a rendición de cuentas consiste en dar a todas las unidades elegibles la misma oportunidad de ser la primera, segunda, tercera, etc. en acceder al programa, lo que implica una implementación aleatoria del programa a lo largo del tiempo.

En los casos en que los recursos son limitados, es decir, en los que nunca habrá suficientes recursos para alcanzar la plena implementación (celdas A1 y A2, y B1 y B2), puede producirse muy rápidamente un exceso de demanda de aquellos recursos. Un sorteo para decidir quién entra en el programa puede ser un enfoque viable para decidir a quién asignar beneficios entre unidades igualmente elegibles. En este caso, cada unidad elegible tiene la misma oportunidad de beneficiarse del programa. Un sorteo es un ejemplo de regla operativa equitativa, transparente y sujeta a rendición de cuentas para asignar los beneficios del programa entre las unidades elegibles.

Otro tipo de programas comprende a los que se implementan a lo largo del tiempo y para los que los administradores pueden clasificar los beneficiarios potenciales en función de la necesidad (celdas A1 y A3). Si los criterios utilizados para priorizar a los beneficiarios son cuantitativos, están disponibles y tienen un umbral de elegibilidad, el programa puede usar un diseño de regresión discontinua.

La otra categoría amplia consiste en programas que tienen la capacidad administrativa para implementarse inmediatamente: es decir, las celdas en

la fila inferior del cuadro. Cuando el programa tiene recursos limitados y no es capaz de clasificar a los beneficiarios (celda B2), podría utilizarse la asignación aleatoria basada en el exceso de demanda. Si el programa tiene suficientes recursos para ampliarse y ningún criterio de elegibilidad (celda B4), la única solución es utilizar variables instrumentales (promoción aleatoria) bajo el supuesto de participación no plena en el programa. Si el programa puede clasificar a los beneficiarios y depende de criterios de elegibilidad, se puede recurrir a la regresión discontinua.

Priorización de los beneficiarios

Las tres preguntas operativas clave guardan relación con el tema fundamental de cómo se seleccionan los beneficiarios, lo cual es crucial para encontrar grupos de comparación válidos. En ocasiones, los grupos de comparación se hallan entre las poblaciones no elegibles, y con mayor frecuencia entre las poblaciones que son elegibles pero que se incorporan al programa más tarde. La manera de priorizar entre los beneficiarios depende en parte de los objetivos del programa ¿Se trata de un programa de jubilaciones para las personas de edad avanzada, un programa de alivio de la pobreza focalizado en los pobres o un programa de inmunización disponible para todos?

Para priorizar entre los beneficiarios sobre la base de la necesidad, el programa debe encontrar un indicador que sea a la vez cuantificable y verificable. En la práctica, la viabilidad de la priorización depende en gran parte de la capacidad del gobierno para medir y clasificar las necesidades. Si el gobierno puede clasificar adecuadamente a los beneficiarios en función de sus necesidades relativas, puede que esté éticamente obligado a implementar el programa de acuerdo con las necesidades. Sin embargo, clasificar en función de la necesidad requiere no solo una medida cuantificable sino también la capacidad y los recursos para medir ese indicador para cada unidad que participa en el programa.

Algunos programas utilizan criterios de selección que, en principio, podrían usarse para clasificar necesidades relativas y determinar la elegibilidad. Por ejemplo, numerosos programas quieren llegar a las personas pobres. Sin embargo, los indicadores de pobreza adecuados que clasifican a los hogares de manera fiable a menudo son difíciles de medir y costosos de recopilar. La recopilación de datos de los ingresos o del consumo de todos los beneficiarios potenciales para clasificarlos según el nivel de pobreza es un proceso complejo y oneroso que, además, sería difícil de verificar. Al contrario, muchos programas utilizan algún tipo de *proxy mean test* para estimar los niveles de pobreza. Se trata de índices de medidas observables sencillas como los activos y las características sociodemográficas (Grosh et al., 2008). Los *proxy mean tests* pueden ayudar a determinar razonablemente bien si un

hogar se sitúa por encima o por debajo de un umbral, pero pueden ser menos precisos en una clasificación detallada de la situación socioeconómica o de las necesidades.

En lugar de enfrentarse al costo y a la complejidad de clasificar a los potenciales beneficiarios individuales, numerosos programas han decidido clasificar en un nivel superior de agregación, como el nivel de la comunidad. Determinar la asignación del programa a un nivel agregado tiene beneficios operativos evidentes, pero a menudo es difícil encontrar indicadores para producir una clasificación de las necesidades en un nivel más agregado.

En los casos en que un programa no puede asignar beneficios de manera fiable sobre la base de la necesidad, ya sea porque no hay indicadores de clasificación cuantificables y verificables, o porque es demasiado caro y propenso a errores, se tienen que usar otros criterios para decidir cómo secuenciar la implementación del programa. Un criterio coherente con la buena gobernanza es la equidad. Una regla equitativa sería dar a todos aquellos que son elegibles la misma oportunidad de ser el primero en tener acceso, y asignar de forma aleatoria un lugar en la secuencia a los beneficiarios potenciales. En la práctica, dadas las dificultades para clasificar las necesidades, una regla de asignación al programa que suele usarse es la asignación aleatoria de los beneficios del programa. También produce un diseño de evaluación aleatoria que puede proveer buena validez interna si se implementa bien, y puede depender de supuestos más débiles en comparación con los otros métodos, como se trata en la sección siguiente.

Una comparación de métodos de evaluación de impacto

Después de estimar qué método de evaluación de impacto es adecuado para las reglas operativas específicas del programa, el equipo de evaluación puede elegir el método que tiene el supuesto más débil y los menores requisitos de datos. El cuadro 11.2 presenta una comparación de los métodos de evaluación de impacto alternativos en términos de los requisitos de datos para implementarlos y los supuestos fundamentales necesarios para interpretar sus resultados como impactos causales de la intervención. Cada fila representa un método diferente. Las primeras dos columnas describen los métodos y las unidades en el grupo de comparación. Las dos últimas columnas recogen los supuestos necesarios para interpretar los resultados como causales, y los datos necesarios para implementar los métodos.

Todos los métodos requieren supuestos, es decir, para ser capaces de interpretar resultados como causales se debe creer que son verdad ciertos hechos que no siempre se pueden verificar empíricamente. En particular,

Cuadro 11.2 Comparación de métodos de evaluación de impacto

Metodología	Descripción	¿Quién está en el grupo de comparación?	Supuesto clave	Datos requeridos
Asignación aleatoria	Las unidades elegibles se asignan de forma aleatoria a un grupo de tratamiento o de comparación. Cada unidad elegible tiene una probabilidad conocida de ser seleccionada. Tiende a generar estimaciones de impacto internamente válidas con los supuestos más débiles.	Las unidades elegibles se asignan aleatoriamente al grupo de comparación.	La aleatorización produce dos grupos estadísticamente idénticos con respecto a las características observables y no observables a lo largo del tiempo en ausencia de la intervención (en la línea de base y a lo largo del seguimiento).	Datos de seguimiento de los resultados en los grupos de tratamiento y comparación; datos de línea de base y otras características para los grupos de tratamiento y comparación con el fin de verificar el equilibrio.
Variables instrumentales (concretamente la promoción aleatoria)	Un instrumento aleatorizado (como una campaña de promoción) induce cambios en la participación en el programa que se evalúa. El método utiliza el cambio en los resultados inducido por el cambio en las tasas de participación para estimar los impactos del programa.	Las unidades que cumplen con los requisitos para participar pero cuya participación se ve afectada por el instrumento (participarían si se exponen al instrumento pero no lo harían en caso contrario).	El instrumento afecta la participación en el programa, pero no afecta directamente los resultados (es decir, el instrumento influye en los resultados solo cambiando la probabilidad de participar en el programa).	Datos de seguimiento de los resultados de todas las unidades; datos sobre la participación efectiva en el programa; datos de los resultados de línea de base y otras características.

Continúa en la página siguiente.

Cuadro 11.2 Comparación de métodos de evaluación de impacto *(continúa)*

Metodología	Descripción	¿Quién está en el grupo de comparación?	Supuesto clave	Datos requeridos
Diseño de regresión discontinua	Las unidades se clasifican a partir de criterios cuantitativos específicos y continuos, como un índice de pobreza. Un umbral determina si una unidad es elegible para participar en un programa. Los resultados de los participantes en una parte del umbral se comparan con los resultados de los no participantes al otro lado del umbral.	Las unidades situadas cerca del umbral, pero que no son elegibles para recibir el programa.	Para identificar impactos no sesgados en el programa para la población cercana al umbral, las unidades que se encuentran inmediatamente por debajo e inmediatamente por encima del umbral son estadísticamente idénticas. Para identificar los impactos no sesgados en el programa para toda la población, la población cercana al umbral debe ser representativa de toda la población.	Datos de seguimiento de los resultados; índice de clasificación y umbral de elegibilidad; datos sobre los resultados de línea de base y otras características.
Diferencias en diferencias	El cambio en el resultado a lo largo del tiempo en un grupo de no participantes se utiliza para estimar cuál habría sido el cambio en los resultados de un grupo de participantes en ausencia de un programa.	Las unidades que no participaron en el programa (por cualquier motivo) y para las cuales se recopilaron datos antes y después del programa.	Si el programa no existía, los resultados de los grupos de participantes y no participantes habrían evolucionado paralelamente a lo largo del tiempo.	Datos de línea de base y de seguimiento de los resultados y otras características tanto para los participantes como para los no participantes.
Paramiento (en particular, pareamiento por puntajes de propensión)	Para cada participante del programa, el método busca la unidad "más similar" en el grupo de no participantes (el pareamiento más estrecho se basa en características observables).	Para cada participante, la unidad no participante que, según las predicciones sobre la base de características observables, tiene la misma probabilidad de haber participado en el programa.	No hay ninguna característica que influya en la participación en el programa más allá de las características observables utilizadas para el pareamiento.	Seguimiento de los datos de los resultados de los participantes y no participantes; datos sobre la participación efectiva en el programa; características de línea de base para llevar a cabo el pareamiento.

Fuente: Adaptado del sitio web de Abdul Latif Jameel Poverty Action Lab (J-PAL).

para cada método, un supuesto clave es que la media del grupo de comparación de la que depende el método sea una estimación válida del contrafactual. En cada uno de los capítulos sobre los métodos, que se presentan en la segunda parte de este volumen, se han expuesto algunas consideraciones sobre cómo probar si un método es válido en un contexto particular. Algunos métodos dependen de supuestos más fuertes que otros.

Ceteris paribus, el método preferido es el que mejor se adecua al contexto operativo y el que requiere los supuestos más débiles y la menor cantidad de datos. Estos criterios explican por qué los investigadores consideran la asignación aleatoria la regla de oro, y por qué a menudo es el método preferido. La asignación aleatoria se adecua a numerosos contextos operativos y tiende a generar estimaciones de impacto internamente válidas con los supuestos más débiles. Cuando se implementa de manera adecuada, genera comparabilidad entre los grupos de tratamiento y comparación en características observables y no observables. Además, la asignación aleatoria tiende a requerir muestras más pequeñas que las necesarias para implementar métodos cuasi-experimentales (véase el debate en el capítulo 15). Dado que la asignación aleatoria es relativamente intuitiva, el método también facilita la comunicación de resultados a los responsables de las políticas.

Concepto clave

El método de evaluación de impacto preferido es aquel que se adecua mejor al contexto operativo, requiere los supuestos más débiles y la menor cantidad de datos.

Puede que los métodos cuasi-experimentales sean más adecuados en algunos contextos operativos, pero requieren más supuestos con el fin de que el grupo de comparación provea una estimación válida del contrafactual. Por ejemplo, el método de diferencias en diferencias depende del supuesto de que los cambios en los resultados en el grupo de comparación proporcionen una estimación válida del cambio del contrafactual en los resultados del grupo de tratamiento. Este supuesto de que los resultados en los grupos de tratamiento y comparación evolucionan paralelamente a lo largo del tiempo no es siempre posible de probar sin múltiples rondas de datos antes de la intervención. La regresión discontinua depende de la comparabilidad de las unidades justo por encima y justo por debajo del umbral de elegibilidad. El pareamiento tiene los supuestos más fuertes de todos los métodos, y esencialmente descarta cualquier característica no observable entre los participantes del programa y los no participantes. En general, cuanto más fuertes sean los supuestos, mayor será el riesgo de que no se cumplan en la práctica.

Un plan de respaldo para la evaluación

A veces las cosas no salen exactamente como estaban planificadas, incluso con el mejor diseño de evaluación de impacto y las mejores intenciones. Por ejemplo, en un programa de capacitación laboral, la agencia ejecutora planeó seleccionar a los participantes de forma aleatoria entre el

conjunto de postulantes, sobre la base del exceso de solicitudes previsto en el programa. Dado que el desempleo entre la población focalizada era alto, se anticipó que el número de postulantes al programa de capacitación laboral sería mucho mayor que el número de plazas disponibles. Desafortunadamente, la publicidad para el programa no fue tan efectiva como se esperaba y, al final, el número de postulantes se situó justo por debajo del número de plazas de capacitación disponibles. Sin un exceso de solicitudes del cual extraer un grupo de comparación, y sin plan de respaldo, el intento inicial para evaluar el programa tuvo que dejarse de lado por completo. Este tipo de situación es habitual, como sucede con los cambios no anticipados en el contexto operativo o político de un programa. Por lo tanto, es útil tener un plan de respaldo en caso de que la primera opción de metodología no funcione.

Planificar el uso de varios métodos de evaluación de impacto también es una buena práctica desde un punto de vista metodológico. Si se plantean dudas acerca de si uno de los métodos puede tener sesgos, se podrán verificar los resultados comparándolos con el otro método. Cuando se implementa un programa mediante asignación aleatoria en fases, el grupo de comparación se incorporará eventualmente al programa. Aquello limita el tiempo durante el cual el grupo de comparación está disponible para la evaluación. Sin embargo, si además del diseño de asignación aleatoria también se implementa un diseño de promoción aleatoria, habrá un grupo de comparación disponible para toda la duración del programa. Antes de que se incorpore el grupo final de la implementación, existirán dos grupos de comparación alternativos (de la asignación aleatoria y de la promoción aleatoria) aunque en el plazo más largo solo quedará el grupo de comparación de la promoción aleatoria.

Cómo encontrar la unidad de intervención más pequeña factible

En general, las reglas operativas también determinan el nivel en que se puede asignar una intervención, algo que se relaciona con la manera en que se implementa el programa. Por ejemplo, si se pone en marcha un programa de salud a nivel de distrito, todas las comunidades del distrito o recibirían el programa (como grupo) o no lo recibirían. Algunos programas se pueden implementar de manera eficiente a nivel individual o de los hogares, mientras que otros deben aplicarse a nivel de la comunidad o a un nivel administrativo superior. Incluso si un programa se puede asignar e implementar a nivel individual, el equipo de evaluación quizá prefiera un nivel superior de agregación con el fin de mitigar los efectos potenciales de derrame; es decir,

los efectos indirectos de las unidades que participan en las unidades que no participan (véase una descripción en el capítulo 9).

Implementar una intervención a un mayor nivel puede ser problemático para la evaluación, por dos motivos. En primer lugar, las evaluaciones de las intervenciones asignadas e implementadas a niveles superiores, como la comunidad o el distrito administrativo, requieren tamaños de muestra más grandes y serán más costosas, en comparación con las evaluaciones de intervenciones a un nivel más bajo, como el nivel individual o de los hogares. El nivel de intervención es importante porque define la unidad de asignación a los grupos de tratamiento y comparación, y eso tiene implicaciones para el tamaño de la muestra de la evaluación y su costo. En las intervenciones implementadas a niveles superiores, se necesita una muestra más grande para poder detectar el impacto del programa. La idea que subyace a esto se abordará en el capítulo 15, donde se analiza cómo determinar el tamaño de la muestra requerido para una evaluación, y cómo la implementación a niveles más altos crea *clusters* (conglomerados) que incrementan el tamaño requerido de la muestra.

En segundo lugar, a niveles superiores de intervención, es más difícil encontrar un número suficiente de unidades para realizar la evaluación. Sin embargo, la asignación aleatoria solo genera grupos de tratamiento y comparación comparables si se lleva a cabo con un número suficiente de unidades. Por ejemplo, si el nivel de agregación es el de la provincia y el país solo tiene seis provincias, es poco probable que la aleatorización genere equilibrio entre los grupos de tratamiento y comparación. En este caso, imagínese que el diseño de la evaluación asigna tres provincias al grupo de tratamiento y otras tres al grupo de comparación. Es muy poco probable que las provincias del grupo de tratamiento sean similares a las del grupo de comparación, incluso si en cada provincia hay un número grande de hogares. Esto es porque la clave para equilibrar los grupos de tratamiento y comparación es el número de unidades asignadas a los grupos de tratamiento y comparación, no el número de individuos o de hogares de la muestra. Por lo tanto, llevar a cabo una asignación aleatoria en niveles altos de implementación pone en riesgo la validez interna si el número de unidades no es suficiente.

Para evitar los riesgos asociados con la implementación de la intervención en un nivel geográfico administrativo alto, el equipo de evaluación y los administradores del programa tienen que trabajar juntos para encontrar la unidad de intervención más pequeña que sea operacionalmente factible. Diversos factores determinan la unidad de intervención más pequeña factible:

- Las economías de escala y la complejidad administrativa en la implementación del programa.
- La capacidad administrativa para asignar beneficios a nivel individual o de los hogares.

- Preocupaciones potenciales a propósito de posibles tensiones.
- Preocupaciones potenciales acerca de los efectos de derrame y la contaminación del grupo de comparación.

La unidad factible de intervención más pequeña suele depender de las economías de escala y de la complejidad administrativa de realizar el programa. Por ejemplo, un programa de seguro de salud quizá requiera una oficina local para que los beneficiarios presenten reclamos y para pagar a los proveedores. Los costos fijos de la oficina tienen que repartirse entre un gran número de beneficiarios, de modo que puede ser ineficiente implementar el programa a nivel individual y más eficiente si se hace a nivel de la comunidad. Sin embargo, en situaciones con tipos de intervenciones nuevas y no probadas, puede que merezca la pena absorber las ineficiencias de corto plazo e implementar el programa en los distritos administrativos, para asegurar la credibilidad de la evaluación y disminuir los costos de la recopilación de datos.

Algunos administradores de programas sostienen que los programas administrados a nivel local, como los programas de seguro de salud, no tienen las capacidades administrativas para implementar programas a nivel individual. Su preocupación es que sería una carga crear sistemas para prestar diferentes beneficios a diferentes beneficiarios en unidades administrativas locales, y que acaso resulte difícil garantizar que la asignación a los grupos de tratamiento y comparación se implemente siguiendo el diseño. Este último problema es una seria amenaza para una evaluación de impacto, dado que los administradores del programa quizá no puedan poner en marcha el programa de forma consistente siguiendo un diseño de evaluación. En este caso, puede que sea necesaria una implementación a un nivel superior o una simplificación del diseño de evaluación de impacto.

En ocasiones los gobiernos prefieren implementar programas a niveles más agregados, como el de la comunidad, porque les preocupan las tensiones potenciales que surgen cuando los miembros de los grupos de comparación observan que los vecinos en el grupo de tratamiento tienen derecho a los beneficios. Numerosos programas se han llevado a cabo con éxito a nivel individual o de los hogares en las comunidades sin generar tensiones, sobre todo cuando los beneficios se han asignado de manera equitativa, transparente y sujetos a rendición de cuentas. Aun así, tendría que tenerse en cuenta el riesgo de que puedan surgir tensiones en el contexto de una evaluación de impacto específica.

Por último, cuando se asigna un programa y se implementa a nivel muy bajo, como en los hogares o a nivel individual, la contaminación del grupo de comparación puede poner en entredicho la validez interna de la evaluación. Por ejemplo, imagínese que se evalúa el efecto de proporcionar agua corriente en la salud de los hogares. Si se instalan grifos de agua para un

Recuadro 11.1: Programas de transferencias monetarias condicionadas y el nivel mínimo de intervención

La mayoría de las transferencias monetarias condicionadas utiliza a las comunidades como el nivel o la unidad de intervención por motivos administrativos y de diseño de programa, y debido a preocupaciones acerca de los efectos de derrame y de posibles tensiones en la comunidad si el tratamiento se asignara a un nivel más bajo.

Por ejemplo, la evaluación del programa de transferencias monetarias condicionadas Progresa-Oportunidades de México dependía de la implementación del programa a nivel comunitario en las zonas rurales para asignar de forma aleatoria las comunidades a los grupos de tratamiento y comparación. A todos los hogares elegibles de las comunidades de tratamiento se les ofreció la oportunidad de inscribirse en el programa en la primavera de 1998, y a todos los hogares elegibles de las comunidades de comparación se les ofreció la misma oportunidad 18 meses más tarde, en el invierno de 1999. Sin embargo, el equipo de evaluación encontró una correlación considerable en los resultados entre los hogares de las propias comunidades. Por lo tanto, para generar suficiente potencia estadística para la evaluación, necesitaban más hogares en la muestra de lo que habría sido necesario si hubieran sido capaces de asignar los hogares individuales a los grupos de tratamiento y de comparación. Por lo tanto, la imposibilidad de implementar el programa a nivel de los hogares generó requisitos de tamaños más grandes de la muestra y aumentó el costo de la evaluación. Otras dificultades similares afectan a muchos de los programas en el sector de desarrollo humano.

Fuentes: Behrman y Hoddinott (2001); Skoufias y McClafferty (2001).

hogar pero no para su vecino, el hogar de tratamiento bien puede compartir el uso del grifo con un vecino de comparación y, por lo tanto, el hogar vecino no sería una verdadera comparación, dado que se beneficiaría del efecto de derrame.

El recuadro 11.1 ilustra las implicaciones de la selección de un nivel de intervención en el contexto de las transferencias condicionadas. En la práctica, los administradores de programa tienen que optar por la unidad factible de intervención más pequeña que (1) permita contar con un gran número de unidades en la evaluación, (2) mitigue los riesgos para la validez interna, y (3) se ajuste al contexto operativo.

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos con más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).

Referencias bibliográficas

Behrman, J. R. y J. Hoddinott. 2001. "An Evaluation of the Impact of PROGRESA on Preschool Child Height." Documento de discusión Núm. 104. Washington, D.C.: International Food Policy Research Institute.

Grosh, M. E., C. Del Ninno, E. Tesliuc y A. Ouerghi. 2008. *For Protection and Promotion: The Design and Implementation of Effective Safety Nets*. Washington, D.C.: Banco Mundial.

Skoufias, E. y B. McClafferty. 2001. "Is *Progresa* Working? Summary of the Results of an Evaluation by IFPRI." Washington, D.C.: International Food Policy Research Institute.

CAPÍTULO 12

Gestión de una evaluación de impacto

Gestión del equipo, del tiempo y del presupuesto de una evaluación

Una evaluación es una alianza entre un equipo de políticas públicas y un equipo de investigación. Cada grupo depende del otro para el éxito de la evaluación. Juntos, constituyen el equipo de evaluación. La alianza se basa en la comprensión de los roles y responsabilidades respectivas de ambos equipos, un compromiso conjunto con la evaluación y un reconocimiento de lo que motiva a las personas a trabajar en la evaluación. Una alianza efectiva es fundamental para asegurar la credibilidad técnica y el impacto de una evaluación en las políticas públicas.

En este capítulo se describen los elementos de una alianza efectiva, lo cual incluye los roles y responsabilidades de cada equipo. También se analiza cómo funciona la alianza en diferentes etapas del proceso de evaluación y se describen los modelos alternativos de colaboración. El capítulo también aborda cuestiones prácticas de calendario y presupuesto.

Roles y responsabilidades de los equipos de investigación y de políticas públicas

El equipo de investigación: función de la investigación y función de los datos

El equipo de investigación es responsable de la calidad técnica y la integridad científica del trabajo de evaluación. Sus responsabilidades abarcan el diseño de la investigación, la calidad de los datos y el análisis. Los equipos de investigación suelen trabajar con las siguientes personas:

- El *investigador principal* trabaja con los responsables de las políticas y los encargados del programa para: establecer objetivos clave, cuestiones de políticas, indicadores y necesidades de información de la evaluación (a menudo utilizando una teoría del cambio, como una cadena de resultados); decidir cuál es la metodología de evaluación de impacto; desarrollar el plan de evaluación; conformar el equipo de investigación; registrar la evaluación de impacto, obtener aprobaciones de la junta de revisión institucional; preparar un plan de evaluación, incluido un plan detallado de preanálisis; dirigir el análisis de los resultados; y colaborar con el equipo de políticas públicas para divulgar los resultados. El investigador principal debe poder trabajar efectivamente con todo el equipo de evaluación, incluyendo la organización encargada de la recopilación de datos, otros miembros del equipo de investigación y los responsables de las políticas públicas o encargados del programa que utilizan los datos y los resultados de la evaluación. Diversos *investigadores* pueden trabajar con el investigador principal o como co-investigadores principales para liderar o apoyar trabajos analíticos específicos de los elementos, como el muestreo, las evaluaciones cualitativas o el análisis de costo-efectividad.

- Un *gestor de la evaluación o coordinador del trabajo de campo,* que trabaja directamente con el investigador principal en la implementación diaria de la evaluación. Esto significa trabajar con los encargados del programa y los responsables de las políticas públicas en el equipo de políticas públicas y supervisar el trabajo de campo cuando se recopilan los datos primarios. Esta persona es particularmente importante en aquellos casos en que el investigador principal carece de una base local, donde se aplica una evaluación prospectiva que debe ser coordinada estrechamente con la implementación del programa o allí donde se recopilan los datos primarios.

- Un *experto en muestreo*, que orienta el trabajo de cálculo de potencia y muestreo. En el tipo de evaluación de impacto cuantitativa que trata este libro, el experto en muestreo debe llevar a cabo cálculos de potencia para

determinar el tamaño adecuado de la muestra según los indicadores establecidos, seleccionar la muestra, comparar los resultados de la muestra real con los de la muestra diseñada, y ofrecer recomendaciones sobre las implicancias para el análisis en coincidencia con el plan de preanálisis. El investigador principal a menudo lleva a cabo estas funciones directamente o junto con el experto en muestreo.

- Un *equipo de recopilación de datos*, que es el encargado de elaborar los instrumentos de recopilación de datos y los manuales y libros de código correspondientes; debe recopilar, digitar y limpiar los datos, y entregar una base de datos limpia y documentada cuando se requiera una recopilación de datos primarios. El capítulo 16 aborda las fuentes de los datos y los diversos aspectos de la recopilación de los mismos.

El equipo de políticas públicas: función de políticas y función de gestión del programa

El equipo de políticas públicas está formado por responsables de políticas y encargados del programa:

- *Los responsables de las políticas* establecen la agenda de investigación, definen la pregunta fundamental que aborda el estudio, aseguran los recursos adecuados para el trabajo, y aplican los resultados a las políticas. Al comienzo de la evaluación, deben articular con claridad los objetivos tanto del programa como de la evaluación, así como la teoría del cambio y los principales indicadores de interés, lo que incluye el tamaño del efecto mínimo relevante para las políticas de los indicadores de resultado de interés, como se detalla en el capítulo 2. El equipo de políticas públicas tiene conocimiento del diálogo de políticas y de los contactos con las principales partes interesadas con el fin de asegurar que la evaluación se diseñe para ser lo más relevante posible para las políticas, y para garantizar que los interesados y los responsables de la toma de decisiones adecuados participen en momentos clave del proceso de evaluación.

- *Los encargados del programa* trabajan mano a mano con el equipo de investigación para alinear el diseño de evaluación con la implementación del programa. Esto incluye verificar que el diseño de evaluación se base en información precisa de la operación del programa y comprometerse a implementar el programa según lo planificado, en el caso de las evaluaciones prospectivas. Los encargados del programa en el equipo de políticas públicas también suelen gestionar el presupuesto de evaluación y a menudo ayudan al equipo de investigación a supervisar el trabajo de campo en la recopilación de datos.

¿A quién le importa la evaluación y por qué?

Concepto clave

Una alianza efectiva entre el equipo de políticas públicas y el equipo de investigación es crucial para asegurar la credibilidad técnica y el impacto de una evaluación en las políticas.

Desde la perspectiva del equipo de políticas públicas, normalmente el principal interés es saber si el programa o la reforma son efectivos o no, y a qué costo se alcanzaron los resultados. Los encargados locales del programa tendrán interés en asegurar que sus esfuerzos sean valorados y que se les otorgue crédito y visibilidad por su trabajo. Emprender una evaluación de impacto significa un esfuerzo considerable de una amplia gama de partes interesadas, a menudo más allá de los límites de sus responsabilidades diarias. Una buena manera de apreciar estas contribuciones consiste en asegurar que los equipos locales participen activamente en la gama más amplia de actividades de evaluación. Esto se puede conseguir celebrando talleres conjuntos, así como también elaborando publicaciones conjuntas, garantizando la capacitación y el desarrollo de capacidades, y consiguiendo investigadores locales bien situados para que contribuyan de manera adecuada y sirvan como un canal válido entre los equipos de investigación y de políticas.

Las evaluaciones tienen valor en términos de bien público cuando fundamentan una pregunta de interés más allá del interés inmediato del equipo de políticas. Este aspecto suele encerrar un interés primario para los investigadores que analizan preguntas relacionadas con una teoría del cambio. Por ejemplo, los resultados sobre cómo se comportan las personas en ciertas circunstancias o cómo funcionan los canales de transmisión para que los impactos se materialicen permiten extraer lecciones de orden más general y aplicarlas en diferentes contextos. Las evaluaciones de impacto están contribuyendo rápidamente a una base de evidencia global sobre el desempeño de una gama de reformas de programas y políticas, y constituyen repositorios de conocimientos sumamente relevantes para el diseño de programas y políticas. A los donantes y a los institutos relacionados con las políticas a menudo les interesa este valor más amplio de bien público, y cada vez prestan más apoyo financiero para llevar a cabo evaluaciones que contribuyan a esta base de evidencia.

Los investigadores también estarán muy comprometidos con el uso de una metodología de evaluación robusta y defendible, y tendrán que asegurar su participación en el diseño de la evaluación de impacto, en el análisis de los datos y en generar investigación primaria que cumpla con las normas científicas vigentes en las publicaciones académicas. Los equipos de investigación interdisciplinaria tienen el reto añadido de asegurar que exista un entendimiento común entre los miembros del equipo. Diferentes disciplinas, como la medicina y la economía, pueden tener distintos enfoques para registrar los ensayos, reclutar a los sujetos, informar sobre los

resultados o divulgarlos, entre otras cosas. Estas expectativas diversas se aclaran y se entienden mejor al comienzo de una evaluación. Al margen de los diferentes protocolos, se espera que los equipos de investigación sigan normas científicas y principios éticos generalmente aceptados, como se trata en el capítulo 13.

Los diferentes intereses del equipo de políticas y del equipo de investigación pueden crear tensiones que hay que entender y gestionar. Los investigadores tenderán a valorar el rigor técnico en el diseño de una evaluación antes que la viabilidad operativa de la implementación del programa. Puede que a los equipos también les interesen distintas preguntas de la evaluación. Por último, puede que ninguno de los dos equipos tenga interés en publicar resultados matizados o negativos, dado que esto podría reflejarse negativamente en el desempeño del programa para el equipo de políticas públicas y podría tener menos interés académico para el equipo de investigación. Puede que al equipo de políticas también le interese ser selectivo a propósito de qué resultados publicar, mientras que el equipo de investigación valorará la capacidad de publicar toda la gama de resultados.

En el conjunto del equipo de evaluación, es crucial promover una cultura de transparencia y de respeto por la evidencia. A los responsables de las políticas públicas y a los administradores del programa se les debería recompensar por su compromiso con la formulación de políticas basadas en la evidencia. Incluso cuando los resultados no sean favorables, se debería dar crédito a estos actores por haber abogado por la transparencia. De la misma manera, se debería alentar al equipo de investigación a informar sobre los resultados y publicarlos, independientemente de los hallazgos.

La alianza entre el equipo de investigación y el equipo de políticas públicas durante la evaluación

La calidad técnica y el impacto de la evaluación en las políticas públicas dependen de una activa alianza entre el equipo de investigación y el equipo de políticas en cada etapa de la evaluación, a saber: diseño, implementación, análisis y divulgación. El recuadro 12.1 resume algunos principios rectores.

Etapa de diseño. En primer lugar, los responsables de las políticas tienen que estructurar y transmitir con claridad las principales preguntas de la investigación, la correspondiente teoría del cambio y los indicadores clave de interés, así como también asegurar que el equipo de investigación comprenda de forma adecuada estos elementos y los respete. Para asegurar la

Recuadro 12.1: Principios rectores de la participación de los equipos de políticas públicas y de evaluación

- Participar desde el comienzo para maximizar las opciones del diseño de la evaluación y asegurar una asociación efectiva entre los equipos de políticas públicas y de evaluación.
- Tener claro un plan de evaluación de impacto desde el comienzo.
- Comprender los roles, responsabilidades y motivaciones de las diferentes partes interesadas y darles la oportunidad de participar en la evaluación.
- Participar a lo largo de la evaluación para asegurar una alineación adecuada entre la evaluación y la intervención que se evalúa.
- Reconocer y gestionar los riesgos y beneficios, dejando claro qué pueden y no pueden hacer las evaluaciones de impacto.
- Valorar la transparencia y asegurar la objetividad; estar preparados para respetar los resultados, sean buenos o malos.

relevancia de las políticas, el equipo de políticas públicas también tiene que estructurar una estrategia de participación que garantice que se consulte y se informe a las partes interesadas acerca del diseño, de la implementación y de los resultados de la evaluación. Por su parte, los investigadores tienen que aclarar, para el equipo de políticas públicas, las condiciones necesarias para una buena evaluación de impacto. En el caso de las evaluaciones prospectivas, esto significará, primero, verificar con los encargados del programa y los responsables de las políticas del equipo de políticas públicas que las operaciones del programa estén lo suficientemente bien establecidas para asegurar que el programa que se evalúa no cambiará sustancialmente durante la evaluación y, por lo tanto, no volverá irrelevantes los resultados de los objetivos de políticas. El momento ideal para llevar a cabo una evaluación de impacto suele ser aquel en el que un programa ha sido sometido a suficientes pruebas de campo como para afirmar que funciona de la manera prevista –lo cual puede fundamentarse en una buena evaluación de proceso–, pero que no ha sido ampliado, por lo que deja abiertas las opciones para construir contrafactuales adecuados.

En segundo lugar, el equipo de investigación tiene que entender con claridad las reglas operativas del programa, a saber: sus recursos disponibles, sus criterios de elegibilidad para seleccionar a los beneficiarios y el calendario de implementación. El equipo de políticas públicas debería trasmitir claramente estas tres reglas operativas al equipo de investigación, dado que son cruciales para fundamentar las opciones metodológicas disponibles en la evaluación, como se detalla en el capítulo 11.

En tercer lugar, el equipo de investigación debería preparar un plan de evaluación de impacto que contenga a la vez aspectos operativos y de la investigación, y debería compartirlo con los responsables de las políticas para asegurar que la evaluación se centre en las preguntas de interés; que los elementos de colaboración con el equipo de políticas estén definidos, y que el equipo de evaluación sea claro y sencillo acerca de las preguntas que se formulan, y de la naturaleza y del calendario de resultados (véase el recuadro 12.2). También es útil tomar en cuenta los riesgos y las estrategias de mitigación propuestas. Por último, el equipo de investigación debería obtener la aprobación ética de una junta de

Recuadro 12.2: Descripción general de un plan de evaluación de impacto

1. Introducción
2. Descripción de la intervención
3. Objetivos de la evaluación
   3.1 Hipótesis, teoría del cambio, cadena de resultados
   3.2 Preguntas de políticas
   3.3 Indicadores de resultados clave
   3.4 Riesgos
4. Diseño de evaluación
5. Muestreo y datos
   5.1 Estrategia de muestreo
   5.2 Cálculos de potencia
6. Visión general del plan de preanálisis
7. Plan de recopilación de datos
   7.1 Encuesta de línea de base
   7.2 Encuesta(s) de seguimiento
8. Productos por entregar
   8.1 Informe de línea de base
   8.2 Informe de evaluación de impacto
   8.3 Nota informativa sobre políticas
   8.4 Bases de datos, diseño y protocolos de análisis plenamente documentados
9. Plan de divulgación
10. Protocolos éticos sobre protección de sujetos humanos
    10.1 Asegurar el consentimiento informado
    10.2 Obtener aprobación de la junta de revisión institucional
11. Calendario
12. Presupuesto y financiamiento
13. Composición y roles del equipo de evaluación

revisión institucional e inscribir la evaluación en un registro de ensayos (véase el capítulo 13).

Este diálogo durante la etapa de diseño debería arrojar como resultado un compromiso claro y compartido con un plan de evaluación, y con expectativas realistas y responsabilidades mutuamente acordadas de los miembros de los equipos de políticas públicas y de investigación. Este diálogo brinda una oportunidad para que el equipo de investigación aclare tanto el valor de una evaluación de impacto –sobre todo el establecimiento de la causalidad y el carácter generalizable de las conclusiones– como sus limitaciones, por ejemplo: no presentar explicaciones de por qué se obtienen ciertos resultados, el *trade-off* entre el tamaño de la muestra y los cálculos de potencia, o el tiempo requerido para generar ciertos resultados. Este diálogo también proporcionará una oportunidad para que el equipo de políticas especifique preguntas prioritarias y para asegurar que la evaluación esté bien alineada con las preguntas de interés de las políticas públicas.

Etapa de implementación. Los equipos de políticas públicas y de investigación tienen que trabajar juntos para asegurar que la implementación proceda fluidamente y se corrijan los problemas. Por ejemplo, en un ensayo controlado aleatorio, el equipo tiene que acordar la mejor manera de aleatorizar en la práctica. Además, durante esta etapa, la coordinación es especialmente importante para garantizar la fidelidad entre el diseño de evaluación y la implementación del programa.

Etapa de análisis. El análisis que se lleve a cabo debe corresponderse con lo que ha sido delineado en el plan de evaluación y en el más detallado plan de preanálisis. El equipo de investigación debería proporcionar y debatir los resultados con el equipo de políticas públicas en coyunturas clave. Empezando por la línea de base, esto debería incluir una revisión de la calidad de los datos recopilados y la adhesión al plan de evaluación. Esto contribuirá a asegurar que el plan de evaluación previsto en la etapa de diseño siga siendo factible y permita cualquier ajuste necesario que deba introducirse. También es una excelente oportunidad para estudiar qué productos se entregarán en qué etapa del análisis y para ver si la producción de esos resultados progresa adecuadamente con respecto a las necesidades de toma de decisiones del equipo de políticas públicas. Una vez que el equipo de evaluación ha concluido el análisis de impacto, debería presentar y compartir los resultados iniciales con el equipo de políticas para asegurar que se responda a todas las preguntas y preparar la etapa de divulgación.

Etapa de divulgación. En esta etapa, el equipo de políticas públicas tiene que asegurar que los resultados de la evaluación lleguen a las personas adecuadas en el momento adecuado y en el formato adecuado. También es la etapa en que se garantiza que todos los datos de la evaluación estén

documentados de forma apropiada. A menudo, los equipos utilizarán diversas estrategias y vehículos para divulgar los resultados, teniendo presentes los diferentes públicos a los que se dirige, como se señala en el capítulo 14.

Establecer una colaboración

Cómo instituir una alianza

Una evaluación es un equilibrio entre los conocimientos técnicos expertos y la independencia que aporta el equipo de investigación, y la relevancia de las políticas, la orientación estratégica y la coordinación operativa con las que contribuyen los responsables de las políticas y los encargados del programa en el equipo de políticas públicas. Se puede utilizar una gama de modelos para crear e implementar esta alianza entre los equipos de investigación y de políticas públicas.

La elección de la modalidad dependerá del contexto y de los objetivos de la evaluación de impacto, teniendo en cuenta una gama de riesgos. Por un lado, un equipo de investigación totalmente independiente, en colaboración limitada con el equipo de políticas públicas, puede generar una evaluación de impacto desvinculada de las preguntas de interés de políticas, o que use una metodología limitada por la falta de interacciones con los encargados del programa. Por otro lado, un equipo de investigación plenamente integrado con el equipo de políticas públicas puede crear riesgos de conflictos de interés, o conducir a la censura de ciertos resultados si no se aplican los principios de la ciencia abierta (véase el capítulo 13). Además, las evaluaciones a menudo pueden tener múltiples objetivos, entre ellos construir capacidad de evaluación con los organismos del gobierno o sensibilizar a los operadores del programa ante las realidades de sus proyectos al llevarse a cabo en el terreno. Estos objetivos más amplios también pueden determinar parcialmente el modelo que se elegirá.

En general, lo que más importa en la calidad de la evaluación de impacto es si el enfoque de asociación producirá estimaciones no sesgadas de los impactos del programa. Siempre que se respeten los principios éticos de la investigación y la ciencia abierta, la ausencia de sesgo y la objetividad tienden a ser más cruciales para la calidad de la evaluación de impacto que la independencia funcional de los equipos de investigación y de políticas. En la práctica, a menudo se requiere una estrecha colaboración entre ambos para asegurar la elaboración de una estrategia de evaluación de impacto de la más alta calidad.

El modelo de externalización

Para los encargados del programa, siempre atareados gestionando operaciones complejas, a menudo es atractivo contar con un equipo externo encargado

de diseñar e implementar la evaluación de impacto. Los modelos de externalización pueden adoptar diferentes formas. Los administradores de programa a veces intentan externalizar el diseño de la evaluación de impacto, así como la realización de diversas encuestas (normalmente, una encuesta de línea de base y de seguimiento) con una sola entidad en el marco de un contrato amplio. En otros casos, los administradores de programa primero externalizan el diseño y siguen con contratos de diversas fases de la recopilación y del análisis de datos.

La externalización separa en cierta medida el diseño de la implementación de la evaluación de impacto, por lo cual una evaluación se puede considerar más independiente. Sin embargo, externalizar totalmente la evaluación de impacto puede implicar riesgos considerables. Establecer este tipo de relación contractual puede limitar la colaboración entre los equipos de implementación y de investigación (o la entidad contratada para llevar a cabo la evaluación de impacto) del programa.

En algunos casos, se entrega al equipo contratado un conjunto de parámetros del programa previamente definidos, con escaso margen para debatir sobre los planes de diseño y de implementación, o sobre el alcance, para dar forma a la investigación. En otros casos, puede que no estén definidas las reglas del programa y las modalidades de implementación necesarias para diseñar una buena evaluación de impacto. En esos casos, el equipo contratado encargado de dicha evaluación tiene una influencia limitada para asegurar que se definan estos elementos.

En otros casos, puede que el programa ya haya sido diseñado o que la implementación haya comenzado, lo cual puede limitar seriamente las opciones metodológicas de la evaluación. A menudo se pide al equipo contratado que se ajuste ex post a cambios en la implementación del programa, sin participar estrechamente ni recibir información durante la implementación. Estas situaciones pueden conducir a diseños de evaluación subóptimos o a dificultades durante la implementación, dado que el equipo contratado puede tener motivaciones diferentes de las de los investigadores y los responsables de las políticas que han dirigido el diseño de la evaluación.

Por último, la selección y supervisión del equipo contratado puede ser problemática para la unidad de implementación del programa. Se deben tener en cuenta atentamente y desde el comienzo las reglas de adquisiciones para asegurar que la externalización sea eficiente y que no presente conflictos de interés. Ciertas reglas pueden limitar la posibilidad de que un equipo que ha sido contratado para contribuir al diseño de una evaluación de impacto pueda más tarde presentar una oferta para ejecutarla.

Para mitigar estos riesgos, normalmente es preferible que el equipo de políticas públicas ya tenga hecho un diseño de evaluación de impacto, que incluya una estrategia de identificación, indicadores de resultados clave,

cálculos de potencia iniciales y tamaños aproximados de la muestra. Esto contribuirá a orientar las adquisiciones y la contratación, dado que dichos elementos influyen claramente en los presupuestos de la evaluación. El equipo de políticas públicas también debería establecer mecanismos para asegurar una supervisión técnica sólida del diseño y de la ejecución de la evaluación de impacto. Esto podría realizarse a través de un comité de supervisión o mediante una revisión técnica y científica regular de los productos de la evaluación. En su conjunto, estas medidas de mitigación señalan que es probable que el modelo más efectivo no sea totalmente externalizado.

El modelo de alianza

La colaboración entre los equipos de investigación y de políticas públicas no se basa única ni necesariamente en relaciones contractuales. Se pueden establecer alianzas mutuamente beneficiosas cuando los investigadores tienen interés en llevar a cabo investigaciones sobre una pregunta de políticas, y cuando los responsables de políticas y los encargados del programa procuran asegurar que su proyecto cuente con una evaluación de impacto de buena calidad. Los investigadores tienen incentivos para abordar nuevas preguntas que se añadirán a la base de evidencia global, y para ampliar el alcance de la evaluación de impacto y contribuir a que sea más visible. El equipo de investigación puede movilizar parte del financiamiento para la evaluación de impacto si los objetivos de los financiadores están estrechamente alineados con el objeto de investigación de la evaluación.

Otro tipo de modelo integrado que está adquiriendo más relevancia, sobre todo en las instituciones más grandes, como el Banco Mundial y el Banco Interamericano de Desarrollo (BID), utiliza una capacidad de investigación de evaluación de impacto interna para apoyar a los equipos de políticas públicas y del programa.

No obstante, el enfoque de la alianza presenta ciertos riesgos. En determinados momentos, puede que los investigadores procuren incorporar elementos novedosos en la investigación de la evaluación de impacto que quizá no estén totalmente alineados con los objetivos inmediatos de las políticas a nivel local, aunque puedan añadir valor en términos más globales. Por su parte, los responsables de las políticas y los encargados del programa quizá no siempre sepan apreciar el rigor científico necesario para emprender evaluaciones de impacto rigurosas, y quizá tengan una mayor tolerancia que el equipo de investigación a los riesgos potenciales de la evaluación de impacto.

Para mitigar esos riesgos, los objetivos del equipo de investigación y de los equipos de políticas públicas deben estar estrechamente alineados. Por ejemplo, ambos equipos pueden trabajar juntos en un plan de evaluación exhaustivo, definiendo una estrategia detallada, así como los roles y responsabilidades de los respectivos equipos (véase el recuadro 12.2).

El plan de evaluación de impacto también es una instancia para resaltar reglas operativas clave, así como los riesgos operativos potenciales para implementar la evaluación de impacto.

Un compromiso mutuo con una evaluación de impacto recogido en un plan de evaluación claro es esencial para que la alianza funcione fluidamente, aun en ausencia de una relación contractual. Corresponde a las buenas prácticas que este compromiso mutuo adopte la forma de un acuerdo por escrito –por ejemplo, bajo la forma de términos de referencia o un memorando de entendimiento– para establecer los roles, responsabilidades y productos de la evaluación de impacto. Estos aspectos también se pueden incluir en el plan de evaluación de impacto.

El modelo plenamente integrado

Algunas evaluaciones de impacto se implementan en un modelo plenamente integrado donde los equipos de investigación y de implementación del programa son el mismo, y son responsables tanto de la investigación como de las funciones del programa. Los investigadores a veces adoptan este enfoque en los ensayos de eficacia, donde se prueban nuevas intervenciones para la *prueba de concepto*. En este caso, los investigadores generalmente prefieren mantener el control de la implementación para asegurar que el programa se ponga en marcha siguiendo el diseño original lo más estrechamente posible. Si bien los resultados de estas evaluaciones de impacto tienen la mayor capacidad para probar las teorías fundamentales y para establecer si una determinada intervención puede funcionar en circunstancias ideales, el riesgo de este enfoque es que los resultados pueden tener una validez externa limitada.

El recuadro 12.3 presenta algunos ejemplos de diferentes modelos que los equipos de investigación y de políticas públicas pueden utilizar para colaborar.

Recuadro 12.3: Ejemplos de modelos de equipos de investigación y de políticas públicas

Externalización de evaluaciones en la Millenium Challenge Corporation

La Millenium Challenge Corporation (MCC) es una agencia de asistencia de Estados Unidos, creada en 2004, con un fuerte énfasis en la rendición de cuentas y los resultados. Requiere que cada uno de sus programas de desarrollo tenga un plan integral de monitoreo y evaluación, centrándose en las evaluaciones independientes y no sesgadas. Esta focalización llevó a la MCC a desarrollar un modelo en el cual tanto el diseño como la implementación de las evaluaciones están totalmente externalizados con investigadores

Continúa en la página siguiente.

Recuadro 12.3: Ejemplos de modelos de equipos de investigación y de políticas públicas *(continúa)*

externos. Durante los primeros años de operaciones de la MCC, en ocasiones la separación entre el equipo del programa y los investigadores externos contratados para la evaluación creó problemas. Por ejemplo, en Honduras, los investigadores diseñaron un ensayo controlado aleatorio de un programa de capacitación agrícola. Sin embargo, dado que el contrato de implementación se basaba en el desempeño, el implementador tenía un fuerte incentivo para encontrar agricultores con un alto desempeño para el programa. Los agricultores elegibles no fueron asignados de forma aleatoria al programa, lo que invalida el diseño de evaluación. Con la divulgación de las primeras cinco evaluaciones de los programas de capacitación agrícola, la MCC reflexionó sobre las experiencias como esta y llegó a la conclusión de que la colaboración entre los implementadores y los evaluadores es crucial a lo largo del diseño y de la implementación. La organización adaptó su modelo para que sea más selectivo al aplicar las evaluaciones de impacto con el fin de encontrar un equilibrio entre la rendición de cuentas y el aprendizaje.

La integración en Innovations for Poverty Action

En Innovations for Poverty Action (IPA), una organización sin fines de lucro de Estados Unidos, los equipos de investigación y de políticas públicas trabajan juntos desde el comienzo del diseño de la evaluación, y a menudo desde el momento en que se gesta el programa. El modelo de IPA cuenta con una amplia red de oficinas en el terreno, muchas de las cuales están en contacto con organismos del gobierno y otros socios. Desde el momento en que una evaluación se concibe, los investigadores afiliados a IPA, provenientes de una red global de universidades, trabajan con los directores de país en las representaciones relevantes para crear un diseño de evaluación y un plan de implementación. Los directores de país son los encargados de gestionar las relaciones entre los socios y emparejar a los principales investigadores del equipo de investigación con los socios del programa en el equipo de políticas públicas para desarrollar una propuesta para una evaluación. Una vez aprobada una propuesta, contratan al personal de gestión del proyecto para dirigir la recopilación de datos en el terreno, todos trabajando en la oficina local de IPA. La coordinación entre los investigadores y los encargados del programa suele ser estrecha, y en algunos casos las oficinas de IPA también son responsables de implementar la intervención que está siendo evaluada.

Modelos mixtos en el Banco Mundial

En la última década, el Banco Mundial ha ampliado rápidamente el uso de las evaluaciones de impacto prospectivas para estimar los impactos de algunos de los proyectos de desarrollo que financia. Varios grupos –entre ellos Development Impact Evaluation (DIME), Strategic Impact Evaluation Fund (SIEF) y Gender Innovation Lab (GIL)– proporcionan financiamiento y apoyo técnico a las evaluaciones de impacto. Cuando se implementa un proyecto particularmente innovador o donde hay grandes intereses en juego, se definen las actividades de evaluación de impacto, ya sea incorporadas en el proyecto y gestionadas por los gobiernos contrapartes o como actividades independientes manejadas por el Banco

Continúa en la página siguiente.

Recuadro 12.3: Ejemplos de modelos de equipos de investigación y de políticas públicas *(continúa)*

Mundial. Se crea un equipo de evaluación que consiste en un grupo de investigación, el cual abarca una combinación de expertos técnicos y académicos, y un equipo de políticas públicas, que normalmente incluye a los responsables de las políticas públicas, los encargados del programa y los jefes de equipo operativos del proyecto.

Por ejemplo, en Costa de Marfil, una iniciativa conjunta del Banco Mundial, JPAL y el gobierno evaluó un proyecto de empleo y desarrollo de capacidades para jóvenes. Para ello, se creó un equipo de evaluación que comprendía un equipo de investigación compuesto por un jefe de equipo del Banco Mundial, académicos internacionales y expertos locales, y un equipo de políticas públicas que incluía a especialistas de la unidad de implementación del proyecto, el ministerio asociado y el personal del Banco Mundial. El equipo de evaluación identificó los ámbitos de prioridad de la evaluación de impacto. Se creó un ensayo controlado aleatorio prospectivo. El gobierno elaboró preguntas clave y financió la recopilación de datos, en parte contratada con el École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSEA), y en parte llevado a cabo internamente por un equipo especializado en recopilación de datos. El Banco Mundial financiaba las actividades de supervisión técnica e investigación, y dirigía el equipo de evaluación. JPAL contribuía a través de los académicos afiliados. Este modelo ha demostrado ser efectivo para asegurar el rigor científico y la relevancia global, así como la alineación con las prioridades de los responsables de las políticas. Requiere una gestión rigurosa de las asociaciones y una coordinación efectiva entre las diversas partes interesadas en el equipo de evaluación.

Fuentes: Bertrand et al. (2015); IPA (2014); Sturdy, Aquino y Molyneaux (2014).

Elección de un equipo de investigación como socio

Los responsables de las políticas y los encargados del programa también tienen que decidir con quién asociarse. Las preguntas clave son si el equipo de investigación –o partes del mismo– puede ser un equipo local, y qué tipo de ayuda externa se requerirá. La capacidad de investigación varía en gran medida de un país a otro. A menudo se contrata a las empresas internacionales cuando se requieren habilidades concretas, y también pueden asociarse con empresas locales. Las funciones de recopilación de datos generalmente son gestionadas por estas últimas, debido a su profundo conocimiento del contexto y del entorno local. También hay una marcada tendencia mundial a asegurar la plena participación de los investigadores locales en la evaluación de impacto.

A medida que aumenta la capacidad de evaluación, es más habitual que los gobiernos, las empresas privadas y las instituciones multilaterales implementen evaluaciones de impacto en asociación con equipos de investigación locales. La participación de los investigadores locales puede aportar un valor fundamental a la evaluación de impacto gracias a su conocimiento

del contexto local. En algunos países, la autorización de la investigación se concede solo a los equipos que incluyen a investigadores locales. En general, el administrador de la evaluación es el que evalúa la capacidad local y determina quién será responsable de qué aspectos del trabajo de evaluación. Las redes académicas internacionales de evaluación de impacto (como JPAL o IPA), las empresas privadas de investigación o grupos de evaluación de impacto de instituciones internacionales (como DIME y SIEF en el Banco Mundial; o SPD o RES en el BID) pueden ayudar a los equipos de políticas públicas a tomar contacto con investigadores internacionales que tengan los conocimientos técnicos expertos para colaborar en la evaluación de impacto.[1]

Otra pregunta es si trabajar con una empresa privada o con un organismo público. Las empresas privadas o los institutos de investigación pueden ser más fiables para proporcionar resultados de manera oportuna pero, una vez que se ha firmado un contrato, las empresas privadas a menudo están menos dispuestas a incorporar en la evaluación elementos que podrán encarecerla. El equipo de investigación también puede trabajar con instituciones de investigación y universidades, cuya reputación y conocimientos técnicos expertos garantizan que las partes interesadas aceptarán los resultados de la evaluación. Sin embargo, en ocasiones esas instituciones carecen de la experiencia operativa o de la capacidad para ejecutar ciertos aspectos de la evaluación, como la recopilación de datos. Por lo tanto, puede que sea necesario subcontratar algunos aspectos con otro socio. El desarrollo de capacidades en el sector público también puede ser un objetivo y se puede incluir como parte de los términos de referencia de la evaluación de impacto. Cualquiera sea la combinación de contrapartes a la que finalmente se llegue, será esencial efectuar un análisis sólido de las actividades de evaluación de los colaboradores potenciales en el pasado para tomar una decisión bien fundamentada.

Particularmente, cuando se trabaja con un organismo público con múltiples responsabilidades, la capacidad y disponibilidad de un equipo de investigación interno para emprender las actividades de evaluación de impacto tienen que ser estimadas a la luz de otras actividades por las que deben rendir cuentas. Es importante tener conciencia de la carga de trabajo para valorar no solo cómo influirá en la calidad de la evaluación que se lleve a cabo, sino también en el costo de oportunidad de la evaluación con respecto a otras iniciativas de las cuales es responsable el organismo público.

Cómo programar una evaluación en el tiempo

En la primera parte de este volumen se analizaron las ventajas de las evaluaciones prospectivas, diseñadas durante la elaboración del programa. Una planificación previa permite una elección más amplia para generar grupos

de comparación, facilita la recopilación de datos de línea de base y ayuda a las partes interesadas a alcanzar un consenso a propósito de los objetivos del programa y de las preguntas de interés.

Aunque es importante planificar las evaluaciones de forma temprana en la etapa de diseño del proyecto, debería programarse su ejecución para evaluar el programa una vez que alcance la madurez para ser estable. Los proyectos piloto o las reformas incipientes suelen sufrir revisiones, tanto en términos de su contenido como con respecto a cuándo, dónde y por quién serán implementados. Los proveedores del programa necesitarán tiempo para aprender y aplicar de manera consistente las nuevas reglas operativas. Dado que las evaluaciones requieren reglas operativas del programa que sean claras para generar contrafactuales adecuados, es importante ejecutarlas cuando los programas estén bien establecidos.

Otro aspecto clave de la programación en el tiempo es cuánto tiempo se requiere antes de que los resultados se puedan medir. El equilibrio adecuado depende en gran parte del contexto: "Si evaluamos demasiado temprano, existe el riesgo de encontrar un impacto parcial o nulo; si evaluamos demasiado tarde, existe el riesgo de que el programa pierda el apoyo de los donantes y del público o de que se amplíe un programa mal diseñado" (King y Behrman, 2009:56).[2] Para determinar cuándo recopilar los datos de seguimiento, debe tenerse en cuenta una gama de factores que se describen a continuación.

El ciclo del programa, que incluye la duración del programa, el tiempo de implementación y los retrasos potenciales. La evaluación de impacto debe ajustarse al ciclo de implementación del programa; la evaluación no puede impulsar el programa que se evalúa. Por su propia naturaleza, las evaluaciones están sujetas a los plazos del programa y deben alinearse con su duración prevista. También deben adaptarse a los posibles desfases en la implementación cuando los programas tardan en asignar beneficios o se retrasan debido a factores externos.[3] En general, a pesar de que la programación en el tiempo de la evaluación debería incluirse en el proyecto desde el comienzo, los evaluadores deberían estar dispuestos a ser flexibles e introducir modificaciones a medida que se ejecuta el proyecto. Además, deberían adoptarse provisiones para dar seguimiento a las intervenciones, utilizando un sistema de monitoreo de modo que el trabajo de evaluación se fundamente en el progreso real de la intervención.

El tiempo previsto necesario para que el programa influya en los resultados, así como la naturaleza de los resultados de interés. La programación de la recopilación de los datos de seguimiento debe tener en cuenta cuánto tiempo se requiere después de que se ejecute el programa para que los resultados se manifiesten. La cadena de resultados del programa ayuda a identificar los indicadores de resultados y el momento adecuado para medirlos. Algunos

programas (como los de apoyo al ingreso) procuran proporcionar beneficios a corto plazo, mientras que otros (como los de educación básica) procuran tener objetivos a más largo plazo. Además, por su propia naturaleza, ciertos resultados tardan más en manifestarse (como los cambios en la esperanza de vida o la fertilidad a partir de una reforma del sistema de salud) que otros (como los ingresos provenientes de un programa de capacitación).

Por ejemplo, en la evaluación del Fondo de Inversión Social de Bolivia, que contaba con datos de línea de base recopilados en 1993, los datos de seguimiento no fueron recopilados hasta 1998 debido al tiempo que se requería para llevar a cabo las intervenciones (proyectos de agua y saneamiento, centros de salud y escuelas) y para que se manifestaran los efectos en la salud y la educación de la población beneficiaria (Newman et al., 2002). Fue preciso un plazo similar para la evaluación de un proyecto de educación primaria en Pakistán, que utilizó un diseño experimental con encuestas de línea de base y de seguimiento para estimar el impacto de las escuelas comunitarias en los resultados de los alumnos, lo que incluía los logros académicos (King, Orazem y Paterno, 2008). Sin embargo, los datos de seguimiento suelen recopilarse antes de lo que sería recomendable, debido a las presiones para obtener resultados de manera oportuna o por limitaciones del presupuesto y del ciclo del proyecto (McEwan, 2014).

Por lo tanto, la recopilación de datos de seguimiento dependerá del programa bajo estudio, así como también de los indicadores de resultados de interés.

Los datos de seguimiento se pueden recopilar más de una vez, de modo que se puedan tener en cuenta y se puedan contrastar los resultados de corto y mediano plazo. Los datos de seguimiento recogidos durante la implementación del programa quizá no capturen el pleno impacto del mismo si los indicadores se miden demasiado temprano. Aun así, es muy útil documentar los impactos de corto plazo, que también pueden proporcionar información acerca de los resultados previstos a más largo plazo, útiles para producir resultados tempranos de la evaluación de impacto, que pueden estimular el diálogo entre los equipos de investigación y de políticas públicas, y mantener el contacto con la muestra de evaluación de modo de reducir el desgaste de la muestra a lo largo del tiempo.

Las encuestas de seguimiento que miden los resultados de largo plazo después de implementar el programa a menudo producen la evidencia más convincente en lo que se refiere a la efectividad del programa. Por ejemplo, los resultados positivos de las evaluaciones de impacto a largo plazo de los programas de desarrollo infantil temprano (DIT) en Estados Unidos (Currie, 2001; Currie y Thomas, 1995, 2000) y Jamaica (Grantham-McGregor et al., 1994; Gertler et al., 2014) han influido en favor de invertir en intervenciones en la primera infancia.

En ocasiones, los impactos de largo plazo son objetivos explícitos del programa, pero puede que incluso un diseño sólido de evaluación de impacto no pueda resistir al paso del tiempo. Por ejemplo, las unidades del grupo de control pueden comenzar a verse favorecidas por los efectos de derrame de los beneficiarios del programa.

Los equipos pueden recolectar datos de seguimiento más de una vez, de modo que se pueden considerar y contrastar los resultados de corto, mediano y largo plazo.

Ciclos de elaboración de políticas. La programación de una evaluación también debe tener en cuenta cuándo se requiere cierta información para fundamentar las decisiones de políticas y debe sincronizar las actividades de evaluación y de recopilación de datos con momentos clave en la toma de decisiones. La producción de resultados debería programarse para fundamentar los presupuestos, la ampliación del programa u otras decisiones de políticas.

Cómo elaborar un presupuesto para una evaluación

El presupuesto constituye uno de los últimos pasos para hacer operativo el diseño de la evaluación. En esta sección, se analizan algunos datos sobre el costo de una evaluación de impacto, se debate cómo elaborar un presupuesto para una evaluación y se sugieren algunas opciones de financiamiento.

Análisis de los datos de costos

Los cuadros 12.1 y 12.2 proporcionan referencias útiles sobre los costos asociados con la realización de evaluaciones de impacto rigurosas. Contienen datos sobre los costos de las evaluaciones de impacto de diversos proyectos realizados con el apoyo del Fondo Estratégico para la Evaluación de Impacto (SIEF, por sus siglas en inglés), administrado por el Banco Mundial. La muestra del cuadro 12.1 proviene de un estudio exhaustivo de programas que reciben el respaldo de los grupos de investigación sobre DIT y educación en SIEF. La muestra del cuadro 12.2 se seleccionó en función de la disponibilidad de estadísticas actuales sobre presupuestos del conjunto de evaluaciones de impacto financiadas por SIEF.[4]

Los costos directos de las actividades de la evaluación analizados en las muestras que se presentan en los cuadros 12.1 y 12.2 oscilan entre US$130.000 y US$2,78 millones, con un costo promedio cercano a US$1 millón. Aunque estos costos varían en gran medida y pueden parecer elevados en términos

Cuadro 12.1 Costo de las evaluaciones de impacto de una selección de proyectos con apoyo del Banco Mundial

Evaluación de impacto (EI)	País	Costo total de la evaluación de impacto (US$)	Costo total del programa[a] (US$)	Costos de la EI como porcentaje del total de los costos del programa
Proyecto de redes de protección	Burkina Faso	750.000	38.800.000	1,9
Desarrollo de destrezas y empleo para migrantes	China	220.000	50.000.000	0,4
Proyecto de protección social	Colombia	130.000	86.400.000	0,2
Plan piloto de nutrición integrada/ sistema de seguridad social asistencial	Yibuti	480.000	5.000.000	8,8
Programa de inversión en sectores sociales	República Dominicana	600.000	19.400.000	3,1
Incentivos para los maestros basados en el desempeño	Guinea	2.055.000	39.670.000	4,9
Protección social	Jamaica	800.000	40.000.000	2,0
Tratamiento de la desnutrición crónica	Madagascar	651.000	10.000.000	6,1
Centros de cuidado del niño basados en la comunidad (piloto)	Malawi	955.000	1.500.000	38,9
Información y transferencias monetarias no condicionadas	Nepal	984.000	40.000.000	2,4
Asistencia técnica en redes de protección social	Pakistán	2.000.000	60.000.000	3,3
Proyecto de protección social	Panamá	1.000.000	24.000.000	4,2
Primer proyecto de niveles de vida comunitarios	Ruanda	1.000.000	11.000.000	9,1
Intervenciones en información para la rendición de cuentas e incentivos para los profesores	Tanzania	712.000	416.000.000	0,2
Intervenciones en el tamaño de la clase y calidad de los profesores	Uganda	639.000	100.000.000	0,6
Fondo social para el desarrollo 3	Rep. Yemen	2.000.000	15.000.000	13,3
Promedio		**936.000**	**59.798.000**	**6,2**

Fuente: Una muestra de evaluaciones de impacto financiadas por los grupos de investigación sobre desarrollo infantil temprano (DIT) y educación del Fondo Estratégico para la Evaluación de Impacto del Banco Mundial (SIEF).
EI = evaluación de impacto.
a. Los costos totales del programa no incluyen los costos asociados con la evaluación de impacto.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
Construcción de capacidad de los padres para colaborar en la nutrición y la salud infantil	Bangladesh	655.000	2.574 hogares	27	48	5	0	20
Cerrando la brecha del aprendizaje temprano entre los niños romaníes	Bulgaria	702.000	6.000 hogares	74	21	4	1	0
Componente de DIT y nutrición del proyecto de redes de protección de Burkina Faso	Burkina Faso	750.000	4.725 hogares	55	20	3	1	21
Pago a profesores comunitarios	Chad	1.680.000	2978 escuelas	52	14	12	18	4
Intervención en DIT basada en el hogar	Colombia	573.000	1.429 personas	54	36	2	2	7
Plan piloto de nutrición/red de protección social integrada	Yibuti	480.000	1.150 personas	75	0	0	6	18
Supervisión e incentivos para un mayor aprendizaje: el Programa de alto desempeño TCAI	Ghana	498.000	480 escuelas	51	46	3	0	0

Continúa en la página siguiente.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial *(continúa)*

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
Incentivos para los maestros basados en el desempeño	Guinea	2.055.000	420 escuelas	82	9	3	1	4
Apoyo en la prestación de servicios educativos	Haití	436.000	200 escuelas	40	31	17	3	9
Motivación no financiera extrínseca e intrínseca de los profesores	India	448.000	360 escuelas	83	5	11	1	0
Estimulación temprana del niño y rendición de cuentas sociales en la estrategia de desarrollo infantil integrada en India	India	696.000	2.250 personas	49	43	5	3	0
Grupos de autoayuda de mujeres para mejorar la salud, la nutrición, el saneamiento y la seguridad alimentaria	India	844.000	3.000 hogares	52	39	5	1	2
DIT para los pobres	India	1.718.000	2.588 hogares	46	53	1	1	0

Continúa en la página siguiente.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial *(continúa)*

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
Nutrición durante la primera infancia, disponibilidad de proveedores de servicios de salud y resultados vitales como jóvenes adultos	Indonesia	2.490.000	6.743 personas	94	0	2	4	0
Para abordar la desnutrición crónica	Madagascar	651.000	5.000 personas	0	0	66	2	32
Habilidades de los padres, nutrición y prevención integrada de la malaria	Mali	949.000	3.600 personas	58	22	4	5	11
Aumento de la rendición de cuentas en educación a través de asistentes pedagógicos basados en la comunidad	México	268.000	230 escuelas	70	26	3	2	0
Acceso a un modelo de escolarización integral privada	México	420.000	172 personas	45	48	5	1	1
Evaluaciones de impacto aleatorias de diversas intervenciones tempranas en destrezas en alfabetización y lectura	Mozambique	1.762.000	110 escuelas	78	5	4	8	6

Continúa en la página siguiente.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial *(continúa)*

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
DIT integrado y nutrición	Mozambique	1.908.000	6.700 hogares	74	8	5	7	7
Plan piloto de seguro de salud	Nepal	485.000	6.300 hogares	61	33	3	4	0
Información y transferencias no condicionadas en los resultados nutricionales	Nepal	984.000	3.000 personas	57	23	9	1	10
Transferencias monetarias, capacitación de los padres y DIT holístico	Níger	984.000	4.332 hogares	67	18	7	1	7
Entendiendo la dinámica de la información para la rendición de cuentas	Nigeria	1.052.000	120 escuelas	59	25	8	3	6
Programa de reinversión de subsidios y de empoderamiento e iniciativa de salud materno-infantil	Nigeria	2.775.000	5.000 hogares	76	13	6	4	2
Participación de la comunidad en el comité escolar	Pakistán	845.000	287 escuelas	59	15	6	3	18

Continúa en la página siguiente.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial *(continúa)*

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
Mejora de las escuelas privadas para los pobres de zonas rurales	Pakistán	2.124.000	2.000 escuelas	26	25	5	2	42
Selección e impactos motivacionales de contratos basados en el desempeño de los maestros de escuela primaria	Ruanda	797.000	300 escuelas	79	7	3	1	11
Campaña de información en escuelas primarias	Sudáfrica	647.000	200 escuelas	67	24	2	3	4
Probando información en la rendición de cuentas e intervenciones de incentivos para los maestros	Tanzania	712.000	420 escuelas	86	6	7	2	0
Diseño de programas de incentivos efectivos para los maestros	Tanzania	889.000	420 escuelas	85	11	2	2	0
Programa para mujeres con alto riesgo de infección de VIH	Tanzania	1.242.000	3.600 personas	90	7	2	1	0

Continúa en la página siguiente.

Cuadro 12.2 Costos desagregados de una selección de proyectos con apoyo del Banco Mundial *(continúa)*

Evaluación de impacto	País	Costo total[a] (US$)	Tamaño de la muestra	Recopilación de datos (porcentaje)[b]	Personal y consultores (porcentaje)[b]	Viajes (porcentaje)[b]	Divulgación y talleres (porcentaje)[b]	Otros (porcentaje)[b]
Intervenciones relativas al tamaño de la clase y la calidad de los maestros	Uganda	639.000	200 escuelas	82	9	7	2	0
Contrastando la eficiencia de la prestación de servicios educativos en los sectores público y privado	Uganda	737.000	280 escuelas	77	18	3	3	0
Promedio		**1.026.000**		**63**	**21**	**7**	**3**	**7**

Fuente: Una muestra de evaluaciones de impacto financiada por el Fondo Estratégico para la Evaluación de Impacto (SIEF) del Banco Mundial.
a. Los costos estimados no siempre capturan todos los costos de la evaluación, lo que incluye el tiempo del equipo de políticas públicas.
b. Es el porcentaje de los costos totales de la evaluación por categoría. Este costo no incluye los costos del personal local del proyecto, que a menudo participaba intensamente en el diseño y la supervisión de la evaluación, dado que los datos precisos de estos costos no se registran de manera regular.

Concepto clave

Las evaluaciones de impacto suelen constituir solo un pequeño porcentaje de los presupuestos generales del programa. Además, el costo de llevar a cabo una evaluación de impacto debe compararse con los costos de oportunidad de no efectuar una evaluación rigurosa y, por lo tanto, de implementar potencialmente un programa inefectivo.

absolutos, las evaluaciones de impacto suelen constituir solo un pequeño porcentaje de los presupuestos generales del programa. Además, el costo de llevar a cabo una evaluación de impacto debe compararse con los costos de oportunidad de no efectuar una evaluación rigurosa y, por lo tanto, de implementar potencialmente un programa inefectivo. Las evaluaciones permiten a los investigadores y a los responsables de las políticas identificar qué programas o características del programa funcionan, cuáles no funcionan y qué estrategias pueden ser las más efectivas y eficientes para alcanzar los objetivos del programa. En este sentido, los recursos necesarios para implementar una evaluación de impacto constituyen una inversión relativamente pequeña pero importante.

El cuadro 12.2 desagrega los costos de la muestra de evaluaciones de impacto financiadas por el SIEF. Los costos totales de una evaluación incluyen el tiempo del personal del Banco Mundial, los consultores nacionales e internacionales, los viajes, la recopilación de datos y las actividades de divulgación.[5] En estas evaluaciones, como en casi todas en las que no se pueden usar los datos existentes, el costo más importante corresponde a la recopilación de nuevos datos, que equivale, en promedio, al 63% del costo de la evaluación, como se muestra en el cuadro.

Estas cifras reflejan diferentes tamaños y tipos de evaluaciones. El costo relativo de la evaluación de un programa piloto suele ser superior al costo relativo de la evaluación de un programa a nivel nacional o universal. Además, algunas evaluaciones solo requieren una encuesta de seguimiento o pueden usar las fuentes de datos existentes, mientras que otras necesitan llevar a cabo múltiples rondas de recopilación de datos. Los costos de recopilación de datos dependen sobre todo de las capacidades del equipo local, de los recursos disponibles y de la duración del trabajo de campo. Para saber más sobre cómo determinar los costos de una encuesta en un contexto particular, se recomienda al equipo de evaluación que primero entre en contacto con el organismo nacional de estadística y que busque información entre los equipos que hayan llevado a cabo un trabajo de encuestas en el país.

Elaboración de un presupuesto para una evaluación de impacto

Se necesitan muchos recursos para implementar una evaluación de impacto rigurosa, sobre todo cuando se trata de recopilar datos primarios. Las partidas presupuestarias incluyen los honorarios para al menos un investigador principal, un asistente de investigación, un coordinador del trabajo de campo, un experto en muestreo y los encuestadores. También se debe considerar el tiempo del personal del proyecto para proporcionar orientación y apoyo a lo largo de la evaluación. Estos recursos humanos

pueden consistir en investigadores y expertos técnicos de organizaciones internacionales, consultores internacionales o locales y personal del programa local. Los costos de viaje y viáticos también se deben presupuestar. También se deben considerar en la planificación los recursos para la divulgación, con frecuencia en forma de talleres, informes y documentos académicos.

Como se ha señalado, el costo más importante suele ser el relacionado con la recopilación de datos (que incluye la creación y la prueba piloto de la encuesta), los materiales y el equipo para recoger los datos, la capacitación de los encuestadores, sus salarios, los vehículos y el combustible, y las operaciones de digitación de datos. Calcular todos estos costos requiere considerar algunos supuestos, por ejemplo, acerca del tiempo que llevará completar el cuestionario y de la duración de los viajes entre los emplazamientos.

Los costos de una evaluación de impacto pueden repartirse a lo largo de varios ejercicios fiscales. El ejemplo de presupuesto del cuadro 12.3 muestra cómo se pueden desagregar los gastos en cada fase de una evaluación por ejercicio fiscal, con fines de contabilidad y de informes. Una vez más, es probable que las demandas presupuestarias sean mayores durante los años en que se recopilan los datos.

Opciones para financiar las evaluaciones

El financiamiento de una evaluación puede provenir de numerosas fuentes, entre ellas: recursos para el proyecto, presupuestos directos del programa, ayudas a la investigación o financiamiento de los donantes. A menudo, los equipos de evaluación procuran tener una combinación de fuentes para generar los fondos necesarios. A pesar de que el financiamiento para las evaluaciones solía provenir sobre todo de presupuestos para la investigación, un énfasis creciente en la elaboración de políticas basadas en la evidencia ha aumentado el financiamiento proveniente de otras fuentes. En los casos en que es probable que una evaluación zanje una brecha de conocimientos considerable de interés para la comunidad de desarrollo en términos más amplios, y donde se pueda aplicar una evaluación creíble y robusta, se debería instar a los responsables de las políticas a buscar financiamiento externo, debido al bien público que los resultados de la evaluación proporcionarán. Las fuentes de financiamiento son el gobierno, los bancos de desarrollo, las organizaciones multilaterales, los organismos de las Naciones Unidas, las fundaciones, las instituciones filantrópicas, y las organizaciones de investigación y evaluación, como la Iniciativa Internacional para la Evaluación de Impacto.

Cuadro 12.3 Ejemplo de presupuesto para una evaluación de impacto

	Etapa del diseño				Etapa de datos de línea de base			
	Unidad	**Costo por unidad (US$)**	**Número de unidades**	**Costo total (US$)**	**Unidad**	**Costo por unidad (US$)**	**Número de unidades**	**Costo total (US$)**
A. Salarios del personal	Semanas	7.500	2	15.000	Semanas	7.500	2	15.000
B. Honorarios de los consultores				14.250				41.900
Consultor internacional (1)	Días	450	15	6.750	Días	450	0	0
Consultor internacional (2)	Días	350	10	3.500	Días	350	10	3.500
Investigador asistente/Coordinador de trabajo de campo	Días	280	0	0	Días	280	130	36.400
Experto estadístico	Días	400	10	4.000	Días	400	5	2.000
C. Viajes y dietas								
Personal: Vuelos internacionales	Viajes	3.350	1	3.350	Viajes	3.350	1	3.350
Personal: Hotel y viáticos	Días	150	5	750	Días	150	5	750
Personal: Transporte terrestre local	Días	10	5	50	Días	10	5	50
Consultores internacionales: Vuelos internacionales	Viajes	3.500	2	7.000	Viajes	3.500	2	7.000
Consultores internacionales: Hotel y viáticos	Días	150	20	3.000	Días	150	20	3.000
Consultores internacionales: Transporte terrestre local	Días	10	5	50	Días	10	5	50
Coordinador de trabajo de campo: Vuelos internacionales	Viajes		0	0	Viajes	1.350	1	1.350
Coordinador de trabajo de campo: Hotel y viáticos	Días		0	0	Días	150	3	150

Continúa en la página siguiente.

Cuadro 12.3 Ejemplo de presupuesto para una evaluación de impacto *(continúa)*

	Etapa del diseño				Etapa de datos de línea de base			
	Unidad	**Costo por unidad (US$)**	**Número de unidades**	**Costo total (US$)**	**Unidad**	**Costo por unidad (US$)**	**Número de unidades**	**Costo total (US$)**
Coordinador de trabajo de campo: Transporte terrestre local	Días		0	0	Días	10	3	30
D. Recopilación de datos								126.000
Tipo de datos 1: Consentimiento					Escuela	120	100	12.000
Tipo de datos 2: Resultados educativos					Niño/a	14	3.000	42.000
Tipo de datos 3: Resultados de salud					Niño/a	24	3.000	72.000
E. Análisis y divulgación de datos								
Taller(es)								
Divulgación/informes								
Costos totales por etapa	Etapa de diseño			43.450	Etapa de línea de base			198.630

Continúa en la página siguiente.

Cuadro 12.3 Ejemplo de presupuesto para una evaluación de impacto *(continúa)*

	Datos de seguimiento Primera etapa				Datos de seguimiento Segunda etapa			
	Unidad	**Costo por unidad (US$)**	**Número de unidades**	**Costos totales (US$)**	**Unidad**	**Costo unitario (US$)**	**Número de unidades**	**Costo total (US$)**
A. Salarios del personal	Semanas	7.500	22222	15.000	Semanas	7.500	22	15.000
B. Honorarios de los consultores				43.750				38.000
Consultor internacional (1)	Días	450	15	6.750	Días	450	10	4.500
Consultor internacional (2)	Días	350	20	7.000	Días	350	10	3.500
Investigador asistente/Coordinador de trabajo de campo	Días	280	100	28.000	Días	280	100	28.000
Experto estadístico	Días	400	5	2.000	Días	400	5	2.000
C. Viajes y dietas								
Personal: Vuelos internacionales	Viajes	3.350	1	3.350	Viajes	3.350	2	6.700
Personal: Hotel y viáticos	Días	150	10	1.500	Días	150	10	1.500
Personal: Transporte terrestre local	Días	10	5	50	Días	10	5	50
Consultores internacionales: Vuelos internacionales	Viajes	3.500	2	7000	Viajes	3.500	2	7000
Consultores internacionales: Hotel y viáticos	Días	150	20	3.000	Días	150	20	3.000
Consultores internacionales: Transporte terrestre local	Días	10	5	50	Días	10	5	50

Continúa en la página siguiente.

Cuadro 12.3 Ejemplo de presupuesto para una evaluación de impacto *(continúa)*

	Datos de seguimiento Primera etapa				Datos de seguimiento Segunda etapa			
	Unidad	**Costo por unidad (US$)**	**Número de unidades**	**Costos totales (US$)**	**Unidad**	**Costo unitario (US$)**	**Número de unidades**	**Costo total (US$)**
Coordinador de trabajo de campo: Vuelos internacionales	Viajes	1.350	1	1.350	Viajes	1.350	1	1.350
Coordinador de trabajo de campo: Hotel y viáticos	Días	150	3	450	Días	150	3	450
Coordinador de trabajo de campo: Transporte terrestre local	Días	10	3	30	Días	10	3	30
D. Recopilación de datos				126.000				126.000
Tipo de datos 1: Consentimiento	Escuela	120	100	12.000	Escuela	120	100	12.000
Tipo de datos 2: Resultados educativos	Niño/a	14	3.000	42.000	Niño/a	14	3.000	42.000
Tipo de datos 3: Resultados de salud	Niño/a	24	3.000	72.000	Niño/a	24	3.000	72.000
E. Análisis y divulgación de datos								55.000
Taller(es)						20.000	2	40.000
Divulgación/informes						5.000	3	15.000
Costos totales por etapa		Seguimiento (etapa I)		201.530	Seguimiento (etapa II)			254.130
					Total costos evaluación			697.740

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para consultar diversos instrumentos útiles en la planificación e implementación de una evaluación, véase el portal de evaluación del BID (http://www.iadb.org/portalevaluacion), que incluye:
  - ➢ Sección de diseño: Cartas Gantt para ayudar en la programación de las actividades de evaluación de impacto, un instrumento de plantilla de presupuesto para estimar los costos de una evaluación de impacto, y una lista de verificación de actividades centrales que se realizarán.
  - ➢ Sección de implementación: Términos de referencia de la muestra para los investigadores principales, empresas de recopilación de datos y apoyo y supervisión técnica.
- Para directrices e instrumentos útiles en la planificación e implementación de una evaluación, véase el World Bank Impact Evaluation Toolkit (Vermeersch, Rothenbühler y Sturdy, 2012), que incluye lo siguiente:
  - ➢ Módulo 2: Armado del equipo: términos de referencia de la muestra para los investigadores principales, coordinadores de la evaluación, analistas de datos, investigadores locales, expertos en cálculos de potencia, expertos en calidad de datos, trabajadores locales y otros.
  - ➢ Manuales de campo y programas de capacitación para hogares y centros de salud.
  - ➢ Módulo 3: Diseño: directrices sobre cómo alinear el calendario, la composición del equipo y el presupuesto de su evaluación de impacto, y una plantilla de presupuesto.
  - ➢ Módulo 4: Preparación de la recopilación de datos: información sobre la programación de actividades de recopilación de datos y logro de acuerdos con los interesados sobre la propiedad de los datos; Carta Gantt, presupuesto de recopilación de datos de la muestra.

Notas

1. El acrónimo JPAL corresponde al Abdul Latif Jameel Poverty Action Lab; SPD es la Oficina de Planificación Estratégica y Efectividad en el Desarrollo, y RES es el Departamento de Investigación del BID.
2. Véase King y Behrman (2009) para un análisis detallado de las cuestiones de programación con respecto a la evaluación de programas sociales.
3. "Hay diversos motivos por los que la implementación no es ni inmediata ni perfecta, por qué la duración de la exposición al tratamiento difiere no solo entre diferentes ámbitos del programa sino también en los diferentes beneficiarios últimos, y por qué las diferentes exposiciones pueden generar diferentes estimaciones del impacto de un programa" (King y Behrman, 2009).

4. Si bien los cuadros 12.1 y 12.2 proporcionan referencias útiles, no son representativos de todas las evaluaciones emprendidas por el programa SIEF o el Banco Mundial.
5. En este caso, el costo se calcula como porcentaje de la parte del proyecto financiado por el Banco Mundial.

Referencias bibliográficas

Bertrand, M., B. Crépon, A. Marguerie y P. Premand. 2015. "Cote d'Ivoire Youth Employment and Productivity Impact Evaluation." AEA RCT Registry (9 de octubre). Disponible en https://www.socialscienceregistry.org/trials/763/history/5538.

Currie, J. 2001. "Early Childhood Education Programs." *Journal of Economic Perspectives* 15 (2): 213–38.

Currie, J. y D. Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85 (3): 341–64.

——. 2000. "School Quality and the Longer-Term Effects of Head Start." *Journal of Economic Resources* 35 (4): 755–74.

Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang y S. Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Grantham-McGregor, S., C. Powell, S. Walker y J. Himes. 1994. "The Long-Term Follow-up of Severely Malnourished Children Who Participated in an Intervention Program." *Child Development* 65: 428–93.

IPA (Innovations for Poverty Action). 2014. "Researcher Guidelines: Working with IPA." (1 de septiembre.) Disponible en http://www.poverty-action.org/sites/default/files/researcher_guidelines_version_2.0.pdf.

King, E. M. y J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 (1): 55–82.

King, E. M., P. F. Orazem y E. M. Paterno. 2008. "Promotion with and without Learning: Effects on Student Enrollment and Dropout Behavior." Serie de documentos de trabajo de investigación de políticas Núm. 4722. Washington, D.C.: Banco Mundial.

McEwan, P. J. 2014. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research*. (doi:10.3102/0034654314553127.)

Newman, J., M. Pradhan, L. B. Rawlings, G. Ridder, R. Coa y J. L. Evia. 2002. "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund." *World Bank Economic Review* 16 (2): 241–74.

Sturdy, J., S. Aquino y J. Molyneaux. 2014. "Learning from Evaluation at the Millennium Challenge Corporation." *Journal of Development Effectiveness* 6 (4): 436–50.

Vermeersch, C., E. Rothenbühler y J. Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. Washington, D.C.: Banco Mundial. Disponible en http://www.worldbank.org/health/impactevaluationtoolkit.

CAPÍTULO 13

La ética y la ciencia en la evaluación de impacto

La gestión de evaluaciones éticas y creíbles

La ética de la evaluación se centra en proteger a los individuos o sujetos humanos que participan en la evaluación, mientras que la transparencia de los métodos contribuye a asegurar que los resultados de la evaluación no estén sesgados, sean fiables y creíbles, y aporten a un acervo más amplio de conocimientos.

Los responsables de las políticas y los investigadores tienen un interés conjunto y una responsabilidad compartida en asegurar que la evaluación sea ética y que sus resultados no estén sesgados, sean fiables y creíbles. Lo contrario puede significar la invalidación de la evaluación y podría generar problemas más allá del alcance mismo de esta. Imagínese una evaluación de impacto que pone en peligro a un grupo de personas al divulgar datos personales, o una evaluación que utiliza un mecanismo de asignación de programa que es injusto porque excluye a las familias más necesitadas. O piénsese en una evaluación que demuestra que un programa es sumamente exitoso pero no divulga ningún dato para apoyar esa afirmación. Cualquiera de estos casos podría producir indignación pública: quejas en los medios de comunicación, en los tribunales o en otras instancias, y sería fuente de malestar para los responsables de las políticas públicas y los investigadores. La crítica de la evaluación podría llegar hasta el programa mismo e incluso

atentar contra su implementación. La fiabilidad y la completitud de los resultados de evaluación también son muy importantes: cuando las evaluaciones producen estimaciones sesgadas o parciales del impacto de los programas, los responsables de las políticas se verán limitados en su capacidad de adoptar una decisión plenamente fundamentada.

Aunque las evaluaciones de impacto estén vinculadas a programas y proyectos públicos, también constituyen una labor de investigación y, por lo tanto, se llevan a cabo en el dominio de las ciencias sociales. De la misma manera, el equipo evaluador debe respetar una serie de principios y reglas de las ciencias sociales para garantizar que la evaluación sea ética y transparente en sus métodos y resultados.

La ética de llevar a cabo evaluaciones de impacto

Cuando una evaluación de impacto asigna los sujetos a grupos de tratamiento y comparación y recopila datos de análisis acerca de ellos, el equipo de evaluación asume la responsabilidad de minimizar en la mayor medida posible cualquier riesgo de que los individuos resulten perjudicados, así como de asegurar que los individuos que participen en la evaluación lo hagan a través de un consentimiento informado.

La ética en la asignación de grupos de tratamiento y comparación

Como sucede con el juramento hipocrático de la profesión médica, un primer principio de la ética de la evaluación debería ser no causar perjuicios. La principal preocupación es que la intervención del programa que se evalúa pueda damnificar a los individuos, ya sea directa o indirectamente. Por ejemplo, un proyecto de rehabilitación de carreteras podría desplazar a los hogares que residen a lo largo de determinadas partes de una carretera. Un proyecto de alfabetización que no tiene en cuenta el uso de las lenguas nativas podría vulnerar a las comunidades indígenas. Numerosos gobiernos y donantes internacionales que financian proyectos de desarrollo utilizan un marco de salvaguardias para evitar y mitigar este tipo de riesgos. Aunque los encargados del programa tienen la responsabilidad fundamental de aplicar medidas de salvaguardias en los proyectos, el equipo de evaluación debería estar vigilante para verificar que el proyecto cumpla con estos marcos requeridos.

Existe otra preocupación a propósito del daño que puede surgir de privar a beneficiarios potenciales de una intervención. Un principio fundamental es que los grupos no deberían ser excluidos de una intervención que

se sabe que es beneficiosa, únicamente por el propósito de llevar a cabo una evaluación. Las evaluaciones solo deberían realizarse en casos en que el equipo de evaluación ignore si una intervención es beneficiosa en el contexto particular en que se evalúa. Además, si una evaluación demuestra que un programa es costo-efectivo, los financiadores del mismo –ya sean gobiernos, donantes u organizaciones no gubernamentales (ONG)– deberían hacer un esfuerzo razonable para ampliar el programa con el fin de incluir a los grupos de comparación una vez que haya finalizado la evaluación de impacto.

Concepto clave

No se debería excluir a un grupo de una intervención que se sabe que es beneficiosa únicamente para los fines de una evaluación.

Un principio relacionado que se impulsa en este libro es que las evaluaciones no deberían dictar cómo se asignan los programas; al contrario, deberían ajustarse a las reglas de asignación del programa en la medida en que estas sean claras e imparciales. La evaluación también puede contribuir a (re)definir las reglas cuando estas no existen o cuando no son justas. Siguiendo este procedimiento, se contribuirá a asegurar que las preocupaciones éticas no emanen tanto de la propia evaluación de impacto como de la ética de las reglas utilizadas para elegir a los beneficiarios del programa. Aun así, la asignación de grupos de tratamiento y comparación puede suscitar inquietudes acerca de la ética de negar los beneficios del programa a los beneficiarios elegibles. Esto es lo que ocurre en particular con la asignación aleatoria de los beneficios del programa. En la segunda parte y en el capítulo 11, se ha puesto de relieve que la asignación aleatoria es un método que se puede aplicar en contextos operativos específicos. Concretamente, el hecho de que la mayoría de los programas funcionen con recursos financieros y administrativos limitados hace imposible llegar a todos los beneficiarios elegibles al unísono. Esto tiene que ver con preocupaciones éticas, dado que el programa mismo debe desarrollar reglas de asignación e imponer algún tipo de selección, incluso sin la existencia de una evaluación de impacto. Desde una perspectiva ética, hay buenos argumentos para que todos aquellos que son igualmente elegibles para participar en un programa tengan iguales probabilidades de ser destinatarios del mismo. La asignación aleatoria cumple este requisito. En otros contextos operativos en los que un programa se va a desarrollar por fases, la implementación se puede basar en la selección aleatoria del orden en que los beneficiarios o grupos de beneficiarios igualmente elegibles serán objeto del programa. En dichos casos, esto dará a cada beneficiario elegible la misma probabilidad de ser el primero en recibir el programa. Así, los beneficiarios que ingresan posteriormente en el programa pueden ser tomados como grupo de comparación para los primeros beneficiarios, generándose de este modo un sólido diseño de evaluación y un método transparente y equitativo para asignar los recursos escasos.

Por último, también puede manifestarse una inquietud ética a propósito de no llevar a cabo una evaluación cuando los programas invierten recursos considerables en intervenciones cuya efectividad se desconoce. En este contexto, la propia falta de evaluación podría verse como no ética porque podría perpetuar programas despilfarradores que no benefician a la población, mientras que los fondos podrían ser mejor gastados en intervenciones más efectivas. La información acerca de la efectividad del programa que producen las evaluaciones de impacto puede contribuir a una inversión más ética y efectiva de los recursos públicos.

La protección de sujetos humanos durante la recopilación, el procesamiento y el almacenamiento de datos

Una segunda instancia en la cual los sujetos podrían verse perjudicados es durante la recopilación, el procesamiento y el almacenamiento de datos. Los hogares, los maestros, médicos, administradores y otras personas que responden a cuestionarios o proporcionan datos a través de otros medios podrían verse perjudicados si la información que proporcionan se divulga públicamente sin suficientes salvaguardias para proteger su anonimato. El perjuicio podría afectar a los propios individuos o a una organización a la que pertenecen. He aquí unos cuantos ejemplos:

- Mientras se lleva a cabo una encuesta, una mujer comparte información acerca de sus prácticas de planificación familiar y su marido (que no está a favor de la planificación familiar) escucha su conversación con el encuestador.
- La privacidad de los hogares se ve violentada (y su seguridad puesta en peligro) cuando un individuo consigue utilizar los datos de una encuesta que fueron publicados en Internet para identificar el ingreso y los activos de familias específicas.
- Un estudio utiliza encuestadores no calificados para realizar pruebas biomédicas, como extracciones de sangre.
- Un encuestado solicita que se le elimine de un estudio a medio camino de la entrevista, pero el encuestador lo insta a acabar de contestar las preguntas.
- Los datos de la encuesta se emplean para identificar a organizaciones comunitarias que se oponen a ciertas políticas de los gobiernos, con el fin de tomar represalias contra ellas.

Frente a riesgos como estos, compete a los investigadores principales y a otros miembros del equipo de investigación salvaguardar los derechos y el

bienestar de los sujetos humanos que participan en la evaluación de impacto, de conformidad con el código ético y la legislación nacional adecuada y con las directrices internacionales.[1] La Organización Mundial de la Salud (OMS) recomienda los siguientes criterios básicos para evaluar los proyectos de investigación con sujetos humanos:

- Los derechos y el bienestar de los sujetos que participan en la evaluación de impacto deberían ser protegidos de forma adecuada.
- Los investigadores deberían obtener un consentimiento informado de los participantes.
- El equilibrio entre riesgo y beneficios potenciales implicados deberían ser valorados y declarados aceptables por un panel de expertos independientes.
- Deberían cumplirse todos los requisitos nacionales especiales.

El Informe Belmont "Principios éticos y pautas para la protección de los sujetos humanos en la investigación" identifica tres preceptos que constituyen el fundamento de la conducta ética de la investigación con sujetos humanos:

- *El respeto por las personas.* ¿Cómo obtendrán los investigadores el consentimiento informado de los sujetos de su investigación?
- *Beneficencia.* ¿Cómo asegurarán los investigadores que la investigación (1) no cause perjuicios y (2) maximice los beneficios potenciales y minimice el daño potencial?
- *Justicia.* ¿Cómo asegurarán los investigadores que los beneficios y cargas de la investigación sean compartidos de forma imparcial y equitativa?

Como elemento clave de su deber de proteger a los sujetos humanos, el investigador principal debería presentar la investigación y los protocolos de recopilación de datos para que sean analizados y aprobados por una Junta de Revisión Institucional (JRI), también conocida como Comité Ético Independiente o Junta de Revisión Ética. La JRI es un comité que ha sido formalmente nombrado para revisar, aprobar y monitorear la investigación biomédica y conductual que trabaja con sujetos humanos. Tanto antes de que comience el estudio como durante su implementación, la JRI revisa los protocolos de investigación y materiales relacionados con el fin de evaluar la ética de la investigación y sus métodos. En el contexto de las evaluaciones de impacto, el análisis de la JRI es particularmente importante cuando el estudio requiere la recopilación de datos de los hogares y de las personas. Concretamente, el estudio de la JRI verifica si los participantes son capaces

de tomar la decisión de participar de las actividades de recopilación de datos, y si su elección estará plenamente fundamentada y será voluntaria. Por último, la JRI analiza si hay algún motivo para creer que la seguridad de los participantes podría estar en riesgo.

El investigador principal tiene la responsabilidad de identificar todas las instituciones que deberían revisar y aprobar el estudio. Numerosos países cuentan con una junta de revisión ética nacional y la mayoría de las universidades tiene una JRI. Normalmente, al equipo se le pedirá que obtenga la aprobación ética de la junta de revisión ética nacional correspondiente del país y de las JRI de cualquier universidad con la que los investigadores tengan alguna filiación. Puede que haya instancias concretas en que las evaluaciones de impacto se lleven a cabo en países que no tienen una JRI nacional o con investigadores cuyas instituciones carecen de dicha junta. En esos casos, el investigador principal debería contratar una JRI (posiblemente comercial) con una tercera parte. El proceso de análisis y aprobación puede tardar entre dos y tres meses, aunque el plazo varía en función de la frecuencia con que se reúne el Comité de la JRI. Los equipos de políticas públicas y de investigación deberían coordinar las presentaciones a la JRI y las actividades de recopilación de datos, de modo que puedan obtener todas las aprobaciones requeridas antes de iniciar la recopilación de datos que involucra a sujetos humanos.

La revisión de la JRI es una condición necesaria pero insuficiente para asegurar la protección de los sujetos humanos. Las JRI pueden variar en gran medida en términos de capacidad y experiencia con los experimentos en ciencias sociales, así como en la focalización de su estudio. Las JRI, sobre todo si están situadas lejos del lugar donde se lleva a cabo la evaluación, pueden no conocer lo suficiente las circunstancias locales para ser capaces de identificar amenazas contextuales de los sujetos humanos. Puede que pongan un énfasis excesivo en la redacción de los cuestionarios y de los formularios de consentimiento. O puede que tengan experiencia en un ámbito más focalizado, como los experimentos médicos, cuyas normas son bastante diferentes de las de los experimentos sociales, en términos de los riesgos para los sujetos humanos. El pensar en la protección de los sujetos humanos no es algo que finaliza una vez que se obtiene la aprobación de una JRI; más bien, debería verse como un punto de partida para asegurar que la evaluación sea ética.

Concepto clave

Una junta de revisión institucional (JRI) es un comité nombrado para estudiar, aprobar y monitorear la investigación con sujetos humanos.

Las juntas de revisión institucional suelen requerir la siguiente información, que debe presentarse para ser analizada:

Evidencia de capacitación. Numerosas JRI (así como muchas directrices éticas nacionales) requieren que el equipo de investigación esté capacitado en la protección de sujetos humanos, aunque las modalidades varían de un país a otro. Al final de este capítulo, en la sección “Otros recursos”, se exponen diversas opciones de capacitación.

El protocolo de investigación. El protocolo de investigación incluye elementos centrales normalmente definidos en el plan de evaluación –sobre todo, la finalidad del estudio y los objetivos de la evaluación, las preguntas centrales de las políticas públicas y la metodología de evaluación propuesta–, así como también la descripción de cómo el equipo de investigación asegurará la protección de los sujetos humanos. Como tal, es un documento importante en la documentación de una evaluación. El protocolo de investigación suele incluir los siguientes elementos en relación con el tratamiento de sujetos humanos: el criterio para seleccionar a los participantes del estudio (sujetos), la metodología y los protocolos aplicados para la protección de sujetos vulnerables, los procedimientos para asegurar que los sujetos sean conscientes de los riesgos y beneficios de participar en el estudio, y los procedimientos utilizados para garantizar el anonimato. La empresa encuestadora debería emplear el protocolo de investigación para orientar los seguimientos del trabajo de campo. En el sitio web de la OMS y en el Kit de Herramientas de Evaluación de Impacto[2] se presenta más información sobre el contenido del protocolo de investigación.

Procedimientos para solicitar y documentar el consentimiento informado. El consentimiento informado es una piedra angular de la protección de los derechos de los sujetos humanos en cualquier estudio. Exige que los encuestados comprendan claramente la finalidad, los procedimientos, los riesgos y beneficios de la recopilación de datos en que se les pide participar. Por defecto, el consentimiento informado de un encuestado adulto requiere un documento por escrito que incluya una sección sobre los métodos utilizados para proteger la confidencialidad del encuestado, una sección sobre el derecho del encuestado a rechazar o cesar su participación en cualquier momento, una explicación sobre riesgos y beneficios potenciales, información de contacto en caso de que el encuestado quiera contactar al equipo de recopilación de datos, y espacio para que los encuestados registren su consentimiento formal por escrito para participar en la recopilación de datos mediante una firma. En ocasiones, los participantes del estudio no son capaces de tomar la decisión de participar. Por ejemplo, a los niños se les suele considerar no capaces de tomar esta decisión. Por lo tanto, al contrario de los adultos, los menores no pueden expresar su consentimiento para participar en una encuesta; pueden acceder a participar si cuentan con un permiso por escrito de sus padres o tutores. Si bien los pasos descritos constituyen los procedimientos por defecto, numerosas evaluaciones de impacto requieren que su JRI les exima del requisito de obtener un consentimiento formal por escrito de los encuestados. Por ejemplo, cuando trabajan con una población analfabeta, a menudo se exime a los potenciales adultos encuestados del consentimiento formal por escrito, que se sustituye por un consentimiento verbal documentado.[3]

Concepto clave

El *consentimiento informado* es una piedra angular de la protección de los sujetos humanos. Exige que los encuestados tengan una clara comprensión de la finalidad, los procedimientos, los riesgos y los beneficios de la recopilación de datos en la que se les pide participar.

Procedimientos para proteger la confidencialidad del encuestado. La protección de la confidencialidad del encuestado es crucial cuando se trata de almacenar y divulgar datos públicamente. Toda la información proporcionada a lo largo de la recopilación de datos debería ser anónima para proteger la identidad de los encuestados. A pesar de que los resultados del estudio pueden publicarse, el informe debería redactarse de tal manera que no sea posible identificar un individuo o un hogar. En lo que respecta a la garantía de confidencialidad de los datos, se debería asignar a cada sujeto de la encuesta un único número de identificación encriptado, y se deberían eliminar todos los nombres e identificadores de la base de datos que se divulga públicamente. Los identificadores incluyen cualquier variable que permita el reconocimiento de individuos u hogares (como direcciones) o cualquier combinación de variables que haga lo mismo (como una combinación de fecha y lugar de nacimiento, sexo y años de escolarización). En caso de que el equipo de investigación prevea que necesitará los identificadores con el fin de hacer un seguimiento de los entrevistados en una encuesta posterior, puede gestionar una base de datos independiente y guardada en condiciones de seguridad, que vincule los ID individuales encriptados con la información de identificación de los encuestados.[4] Además de encriptar los ID individuales, puede que también sea necesario encriptar las localizaciones e instituciones. Por ejemplo, si los hogares y los individuos están codificados con ID encriptados, pero las localidades están identificadas, puede que sea posible reconocer los hogares a través de las características contenidas en la encuesta. Por ejemplo, puede que una localidad concreta incluya solo un hogar que posee una motocicleta, siete vacas y una peluquería. Cualquiera con acceso a los datos podría ser capaz de localizar el hogar y de esta manera se violaría la confidencialidad.

Garantizar evaluaciones fiables y creíbles mediante la ciencia abierta

Uno de los objetivos fundamentales de la evaluación de impacto consiste en estimar el impacto de un programa en una gama de resultados de interés. En la segunda parte de este capítulo se abordan una serie de métodos para asegurar que los impactos estimados sean robustos. Una evaluación de impacto bien diseñada y bien implementada debería garantizar que los resultados no estén sesgados, sean fiables y creíbles, y que contribuyan a un acervo de conocimiento más amplio. Cuando las evaluaciones no están sesgadas, y son fiables y creíbles y se pueden interpretar dentro de un acervo relevante de conocimiento relacionado, pueden contribuir a que se tomen las decisiones

adecuadas de política pública y se mejoren las vidas de las personas. Sin embargo, en la práctica, hay varios problemas que pueden impedir que este ideal se alcance. En esta sección, se analizará cómo diversas cuestiones científicas de la evaluación de impacto se pueden convertir en asuntos difíciles para los responsables de las políticas, y se presentarán medidas potenciales para evitar o mitigar estos problemas. Estas medidas se suelen agrupar bajo la denominación de *ciencia abierta*, porque su objetivo consiste en lograr que los métodos de investigación sean transparentes.[5] La mayoría de estos problemas tienen que ser manejados por el equipo de investigación, pero el equipo de política que supervisa una evaluación de impacto tiene que ser consciente de los mismos mientras administra las evaluaciones de impacto. En el cuadro 13.1 se resumen los problemas, las implicaciones de política y las posibles soluciones.

Cuadro 13.1 Asegurar información fiable y creíble para las políticas mediante la ciencia abierta

Problemas de la investigación	Implicaciones para las políticas públicas	Soluciones de prevención y mitigación mediante la ciencia abierta
Sesgo de la publicación. Solo se publican los resultados positivos. Las evaluaciones que muestran impactos limitados o nulos no se divulgan ampliamente.	Las decisiones de política se basan en un acervo distorsionado de conocimiento. Los responsables de la política pública tienen escasa información sobre lo que no funciona y siguen probando/adoptando políticas que no tienen impacto alguno.	Registros de ensayos.
Minería de datos. Los datos se fragmentan cada vez más hasta que aparece un resultado positivo en la regresión, o la hipótesis se reajusta a los resultados.	Las decisiones de política para adoptar intervenciones pueden estar basadas en estimaciones positivas no justificadas de los impactos.	Planes de preanálisis.
Pruebas de hipótesis múltiples, análisis de subgrupo. Los investigadores fragmentan cada vez más los datos hasta que encuentran un resultado positivo para algún grupo. Concretamente: (1) las múltiples pruebas conducen a la conclusión de que algunos impactos existen cuando en realidad no existen; o (2) solo se informa sobre los impactos que son significativos.	Las decisiones de política pública para adoptar intervenciones pueden estar basadas en estimaciones positivas no justificadas de los impactos.	Planes de preanálisis y técnicas de ajustes estadísticos especializadas, como las pruebas de índices, la tasa prudente de error relacionada con la familia y el control de la tasa de falsos descubrimientos.[a]

Continúa en la página siguiente.

Cuadro 13.1 Asegurar información fiable y creíble para las políticas mediante la ciencia abierta *(continúa)*

Problemas de la investigación	Implicaciones para las políticas públicas	Soluciones de prevención y mitigación mediante la ciencia abierta
Falta de replicación. No se pueden replicar los resultados porque el protocolo de investigación, los datos y los métodos de análisis no están suficientemente documentados.	La política puede basarse en resultados manipulados (positivos o negativos), dado que los resultados pueden deberse a errores de cálculo.	La documentación y el registro de los datos, incluidos los protocolos de proyecto, los códigos de organización, la publicación de los códigos, y la publicación de datos.
Los errores y las manipulaciones pueden pasar inadvertidos.	Los resultados entre diferentes estudios no se pueden comparar.	Cambios en las políticas de las revistas arbitradas y de financiamiento para requerir documentación acerca de los datos y promover la replicación.
A los investigadores no les interesa replicar los estudios, y a las revistas arbitradas no les interesan los resultados "yo también".[a]	La validez de los resultados en otro contexto no se puede probar.	
No se pueden replicar las intervenciones porque el protocolo de intervención no está suficientemente documentado.	Los responsables de la política pueden ser incapaces de replicar la intervención en un contexto diferente.	

a. Para una introducción básica al problema de las comparaciones múltiples y las correcciones estadísticas potenciales, se recomienda consultar https://en.wikipedia.org/wiki/Multiple_comparisons_problem.

Sesgo en la publicación y registros de pruebas

Normalmente, a los investigadores que trabajan en evaluaciones de impacto les interesa asegurarse de que los resultados de sus evaluaciones sean publicados en revistas arbitradas porque eso contribuye a su carrera profesional. Sin embargo, la mayoría de los resultados que aparecen en estas publicaciones muestran impactos positivos. Por lo tanto, se impone la pregunta de qué sucede con las evaluaciones que tienen resultados negativos o que no pueden mostrar resultados significativos. Los investigadores prácticamente no tienen incentivos para consignar resultados no significativos o someterlos a publicaciones arbitradas porque perciben que hay escaso interés en los resultados y que las revistas rechazarán sus documentos (Franco, Malhotra y Simonovits, 2014). Este sesgo en la publicación suele denominarse "problema del cajón de archivador" porque los resultados permanecen en el "archivador" y no son divulgados ni publicados. Pueden surgir problemas de sesgo de publicación similares en las evaluaciones de impacto de programas específicos. Es más probable que los equipos de política pública, los financiadores y los gobiernos den a conocer y publiciten los resultados positivos de una evaluación de un programa en lugar de difundir resultados negativos o ausencia de resultados. Debido a estas tendencias, es difícil tener un cuadro claro de las intervenciones que no funcionan, dado que los

resultados no suelen estar disponibles y el acervo de evidencia con el que se cuenta está más bien distorsionado. Los responsables de la política pública que tratan de basar sus decisiones en la evidencia disponible quizá no tengan acceso a resultados no publicados; como consecuencia, puede que continúen intentando trabajar con políticas que no han tenido éxito en otros lugares.

Una solución parcial al sesgo de publicación es el registro de las pruebas. Se debería alentar a los equipos de evaluación de impacto a registrar sus pruebas, y en ese sentido el equipo de política pública tiene un importante rol que desempeñar para asegurar que el equipo de investigación registre la evaluación de impacto. El registro de pruebas es muy común (y a menudo requerido) en las ciencias médicas, pero recién comienza a ganar terreno en las ciencias sociales, lo que incluye las evaluaciones de impacto. El registro implica que los investigadores declaren públicamente su intención de llevar a cabo una evaluación antes de hacerlo realmente, dejando asentada información clave acerca de la evaluación en un registro (véase el recuadro 13.1). Como consecuencia, debería ser posible tener una lista completa de las evaluaciones de impacto que se hayan llevado a cabo, tanto si los resultados fueron positivos como si no lo han sido.

Recuadro 13.1: Registro de pruebas en las ciencias sociales

Las evaluaciones de impacto de las políticas públicas normalmente deberían asentarse en los registros de ciencias sociales en lugar de hacerlo en los registros médicos, debido al carácter de la investigación. He aquí unos cuantos ejemplos:

- El registro de la American Economic Association para pruebas aleatorias controladas se puede consultar en http//www.socialscienceregistry.org. En julio de 2015 contenía 417 estudios realizados en 71 países.
- La Iniciativa Internacional para la Evaluación de Impacto (3ie) gestiona el Registry for International Development Impact Evaluations (RIDIE), que se centra en las evaluaciones de impacto relacionadas con el desarrollo de los países de ingresos bajos y medios. En julio de 2015 contaba con alrededor de 64 evaluaciones registradas.
- El Center for Open Science gestiona el Marco de Ciencia Abierta (OSF, por sus siglas en inglés) y tiene un foco ligeramente diferente, pero también puede servir como registro (https://osf.io/). El OSF es un sistema de gestión basado en la nube para proyectos de investigación, que permite crear "fotos instantáneas" de la investigación en cualquier momento del tiempo, con una URL persistente y una marca de fecha. Los investigadores pueden subir su protocolo, investigar hipótesis, datos y códigos en el OSF, y compartir el enlace resultante de la web como prueba de registro.

Los registros constituyen un gran paso hacia adelante para garantizar que el acervo disponible de conocimiento no se distorsione más. Sin embargo, aún persisten muchas dificultades. Por ejemplo, aunque quede claro en un registro que una evaluación se llevó a cabo, puede que no sea tan fácil obtener información acerca de los resultados de la misma. Las evaluaciones de impacto se pueden suspender o pueden no llevarse a cabo. E incluso si se encuentra disponible la falta de resultados de una evaluación, esto a menudo suscita un conjunto adicional de preguntas que complican la interpretación de los resultados: ¿Acaso los investigadores no encontraron resultados porque la evaluación estaba mal diseñada y ejecutada, porque el programa no estaba bien implementado, o porque el programa realmente no tuvo un impacto? Como se verá en el capítulo 16, la recopilación de datos complementarios a través del monitoreo del programa o desde fuentes alternativas de datos puede contribuir a garantizar que los resultados estén bien interpretados.

Minería de datos, pruebas de hipótesis múltiples y análisis de subgrupos

Otro problema potencial relacionado con la evaluación de impacto es la *minería de datos*, la práctica de manipular los datos en búsqueda de resultados positivos. La minería de datos puede manifestarse de diferentes maneras. Por ejemplo, cuando hay datos disponibles, puede que aparezca la tentación de aplicar regresiones sobre los mismos hasta que surja "algo" positivo, y luego reajustar una hipótesis atractiva a aquel resultado. Esto constituye un problema por el siguiente motivo: cuando se aplican pruebas estadísticas para la significancia de los impactos, hay que utilizar un nivel de significancia de, por ejemplo, 5%. Estadísticamente, 1 en 20 pruebas de impacto arrojarán niveles significativos al 5%, aun cuando la distribución subyacente no garantice un impacto (véase el capítulo 15 para un debate sobre los errores de tipo I). Con la minería de datos, ya no se puede garantizar que el resultado de un impacto sea genuino, ni si proviene únicamente de las propiedades estadísticas de la prueba. Este problema está relacionado con la cuestión de las *pruebas de hipótesis múltiples*, a saber: cuando una investigación incluye múltiples hipótesis diferentes, existe una alta probabilidad de que al menos una de ellas se confirme con una prueba positiva únicamente por azar (debido a las propiedades estadísticas de la prueba), y no debido al impacto real. Una situación similar surge en el análisis de subgrupos: cuando la muestra es lo suficientemente grande, los investigadores podrían intentar subdividirla hasta que encuentren un impacto en *algún* subgrupo. Una vez más, no se puede estar seguro de que un resultado de impacto en ese subgrupo sea un resultado genuino, o si proviene únicamente de las propiedades estadísticas de la prueba.

Otro ejemplo de minería de datos tiene lugar cuando la decisión de continuar o suspender la recopilación de datos se vuelve dependiente de un resultado intermedio: por ejemplo, una encuesta de hogares se planificó para un tamaño de muestra de 2.000 hogares y el trabajo de campo ha avanzado hasta los 1.000. Si esta muestra reducida produce un resultado positivo de la evaluación de impacto y se toma la decisión de suspender la recopilación de datos para evitar el riesgo de que más datos puedan cambiar los resultados, esto sería minería de datos. Otros ejemplos son la exclusión de ciertas observaciones o grupos inconvenientes, o el ocultamiento selectivo de resultados que no encajan. Si bien no hay motivos para creer que estas prácticas son generalizadas, unos cuantos casos flagrantes y de alto perfil tienen el potencial para socavar la evaluación de impacto como ciencia. Además, incluso hay casos menores de minería de datos que tienen el potencial de distorsionar el acervo de evidencia utilizado por los responsables de la política pública para decidir qué intervenciones comenzar, continuar o suspender.

Una recomendación habitual para evitar la minería de datos consiste en utilizar un *plan de preanálisis*. Este plan define los métodos de análisis antes de llevar a cabo el análisis de evaluación de impacto, dejando en claro así el foco de la evaluación y reduciendo el potencial para alterar los métodos una vez que haya comenzado el análisis. El plan de preanálisis debería especificar los resultados que se medirán, las variables construidas y utilizadas, los subgrupos para los que se llevará a cabo el análisis y los enfoques analíticos básicos que se utilizarán en la estimación de los impactos. Los planes de preanálisis también deberían incluir las correcciones propuestas por los investigadores en las pruebas de hipótesis múltiples y pruebas de subgrupos, si se requiere. Por ejemplo, probar el impacto de una intervención en educación de seis diferentes puntuaciones de pruebas (matemáticas, inglés, geografía, historia, ciencia, francés) para cinco grupos diferentes de escuelas (grados 1 a 5) y dos sexos (masculino y femenino) arrojaría 60 hipótesis diferentes, una o varias de las cuales están destinadas a tener una prueba significativa solo por azar. Al contrario, el investigador podría proponer calcular uno o más índices que agrupan a los indicadores, para reducir el número de hipótesis y subgrupos.[6]

Aunque un plan de preanálisis puede contribuir a aliviar la preocupación de la minería de datos, también existe la preocupación de que podría eliminar parte de la flexibilidad necesaria en el tipo de análisis que llevan a cabo los investigadores. Por ejemplo, puede que el plan de preanálisis especifique los canales anticipados de impacto de una intervención a través de la cadena de resultados. Sin embargo, una vez que la intervención se implemente en la práctica, de pronto puede surgir un conjunto de factores adicionales no anticipados. Por ejemplo, si un gobierno está pensando en implementar una

nueva manera de pagar a los proveedores de cuidados de salud, uno podría identificar posibles canales de impacto. Sin embargo, sería muy difícil anticipar todos los efectos posibles que esto podría tener. En algunos casos, sería necesario realizar entrevistas cualitativas con los proveedores para entender precisamente cómo se adaptan a los cambios y cómo esto influye en el desempeño. Sería muy difícil incorporar todas estas posibilidades en el plan de preanálisis por adelantado. En ese caso, los investigadores tendrían que trabajar por fuera del plan de preanálisis original, y no ser penalizados por ello. En otras palabras, un plan de preanálisis puede otorgar una credibilidad adicional a las evaluaciones, convirtiéndolas en confirmaciones de una hipótesis en lugar de ser solo investigación exploratoria; sin embargo, los investigadores deberían poder seguir explorando nuevas opciones que se pueden convertir en investigación confirmativa en evaluaciones posteriores.

Falta de replicación

Hay dos tipos de replicaciones importantes para la evaluación de impacto. En primer lugar, en un determinado estudio, los investigadores que no pertenecen al equipo de investigación original deberían poder reproducir los mismos resultados (o al menos muy similares) que los investigadores originales utilizando los mismos datos y análisis. Las replicaciones de un determinado resultado de la evaluación de impacto constituyen una manera de verificar su validez interna y su ausencia de sesgo. Cuando los estudios o los resultados no pueden replicarse debido a la falta de disponibilidad de información sobre la codificación o los datos, existe el riesgo de que los errores y las manipulaciones en el análisis pasen desapercibidos, y de que los resultados imprecisos sigan influyendo en las políticas. Afortunadamente, se están logrando avances sustanciales en términos de hacer disponibles los datos, los códigos y los protocolos. Cada vez más revistas arbitradas de ciencias sociales están comenzando a exigir que esos datos y códigos estén disponibles junto con la publicación de los resultados. Directrices como las de Promoción de la Transparencia y la Apertura, desarrolladas por el Centro para la Ciencia Abierta, están cambiando lentamente las prácticas y los incentivos. Para asegurar que pueda realizarse la replicación, los equipos de evaluación de impacto tienen que hacer disponibles públicamente los datos y asegurar que todos los protocolos (incluido el de aleatorización), las bases de datos, y los códigos de análisis de la evaluación de impacto estén documentados, almacenados en condiciones de seguridad y suficientemente detallados.

En segundo lugar, una vez que se completa una evaluación, debería ser posible que otros responsables de la política pública e investigadores utilicen las intervenciones y protocolos de evaluación originales y los apliquen en un contexto diferente o en un momento distinto para ver si los resultados

se mantienen bajo circunstancias diversas. La falta de replicación de los resultados de la evaluación es un asunto serio para los responsables de la política pública. Por ejemplo, una evaluación muestra que la introducción de computadores en las escuelas tiene resultados sumamente beneficiosos, pero este es el único estudio que produjo esos resultados y otros investigadores no han podido obtener los mismos resultados positivos en posteriores evaluaciones de programas similares. ¿Qué debe hacer un responsable de política pública en ese caso? La falta de replicación de los resultados puede deberse a diferentes causas. En primer lugar, quizá sea difícil llevar a cabo evaluaciones que intenten solo replicar resultados que fueron obtenidos en un estudio anterior: puede que ni a los investigadores ni a los financiadores les interesen los estudios de "yo también". En segundo lugar, aun cuando existan la voluntad y los fondos para replicar los estudios, la replicación no siempre es posible porque puede que los protocolos (incluido el de aleatorización), los datos, y los códigos de análisis del estudio original no estén disponibles ni sean lo suficientemente detallados. Se observa un esfuerzo creciente entre las organizaciones que apoyan las evaluaciones de impacto para alentar replicaciones en diferentes contextos: por ejemplo, desarrollando grupos de estudio sobre temas similares o promoviendo evaluaciones de impacto de multisitios.

Lista de verificación: una evaluación de impacto ética y creíble

Los responsables de la política pública tienen un importante rol que desempeñar para asegurar que se creen las condiciones necesarias para una evaluación de impacto ética y creíble. Concretamente, los responsables de la política pública tienen la responsabilidad fundamental de asegurar que las reglas de asignación del programa sean justas, y pueden pedir al equipo de investigación una rendición de cuentas de la transparencia de los métodos de investigación. A continuación, se sugiere una lista de preguntas de verificación.

- ✓ ¿Es justa la asignación a los grupos de tratamiento y comparación? ¿Hay grupos con necesidades particularmente acuciantes que deberían recibir el programa de todas maneras? ¿Quién será excluido de la evaluación de impacto?
- ✓ ¿El equipo de investigación ha identificado la JRI o el comité de revisión ética nacional pertinente?
- ✓ ¿Permite el calendario de la evaluación de impacto contar con tiempo suficiente para preparar y presentar el protocolo de investigación a la JRI

y obtener consentimiento antes de que comience la recopilación de datos de sujetos humanos?

- ✓ ¿El equipo de investigación presentó el protocolo de investigación y el plan de preanálisis a un registro de pruebas de ciencias sociales?
- ✓ ¿Existe un procedimiento para asegurar que los elementos clave de la intervención estén documentados tal como ocurren en la práctica, y no solo como están planificados?
- ✓ ¿Los responsables de la política pública comprenden que los resultados de la evaluación pueden mostrar que la intervención no fue efectiva, y están de acuerdo en que esos resultados serán publicados y no retenidos?
- ✓ ¿El equipo de evaluación ha identificado la manera en que se divulgarán los datos y los resultados de la evaluación, aun cuando el equipo de investigación no consiga publicar los resultados en una revista arbitrada?

Los principios, los problemas y la lista de verificación definidos en este capítulo pueden contribuir a asegurar que una evaluación de impacto sea creíble y ética.

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Capacitación en investigación con sujetos humanos de los Institutos Nacionales de Salud (National Institutes of Health o NIH) de Estados Unidos.
  - ➢ Los NIH ofrecen una capacitación en línea que, si bien se centra en las ciencias médicas y en Estados Unidos, es muy informativa y se tarda solo una hora en completarla. Véanse los enlaces: http://phrp.nihtraining.com/users/login.php y www.ohsr.od.nih.gov.
- Capacitación en investigación con sujetos humanos a través de la Iniciativa de Capacitación Institucional Colaborativa (CITI) de la Universidad de Miami.
  - ➢ La CITI brinda cursos internacionales en varias lenguas tanto a organizaciones como a individuos, aunque el programa tiene un costo (valor inicial: US$100 por persona). Véase http://www.citiprogram.com.
- Compilación internacional de normas de investigación en seres humanos:
  - ➢ Cada año, el Departamento de Salud y de Servicios Humanos de Estados Unidos publica una compilación de leyes, regulaciones y directrices que rigen la investigación sobre seres humanos. La edición de 2015 incluye 113 países, así como también las normas de diversas organizaciones internacionales y regionales. El documento contiene las juntas de revisión institucional nacionales e internacionales (http://www.hhs.gov/ohrp/international).

- Procedimientos para la protección de sujetos humanos en investigaciones apoyadas por la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID). Véase el enlace http://www.usaid.gov/policy/ads/200/humansub.pdf.
- *Manual de mejores prácticas en la investigación transparente en ciencias sociales*, de Garret Christensen, con la asesoría de Courtney Soderberg (Center for Open Science). Véase el enlace https://github.com/garretchristensen/BestPracticesManual.
  - ➢ Guía de trabajo de las últimas mejores prácticas para la investigación cuantitativa transparente en ciencias sociales. El manual es actualizado de manera regular.
- Directrices de Promoción de la Transparencia y la Apertura (TOP). Véase el enlace http://centerforopenscience.org/top/.
  - ➢ Las directrices se pueden encontrar en el sitio web del Center for Open Science.
- Para enlaces a juntas de revisión independientes reconocidas y servicios independientes de JRI, véase el Portal de Evaluación del Banco Interamericano de Desarrollo (BID): http://www.iadb.org/portalevaluacion.
- Para más información sobre la recopilación de datos, véase el Portal de Evaluación del BID: http://www.iadb.org/portalevaluacion.
  - ➢ Véase la sección sobre recopilación de datos en la sección de protección de sujetos humanos.
  - ➢ Nótese que el enlace de la Association for the Accreditation of Human Research Protection Programs (AAHRPP) ofrece capacitación y certificación para las JRI. Se puede encontrar una lista de las organizaciones acreditadas en su sitio web.
- Para directrices sobre la protección de los participantes, véase el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial, Módulo 4 (http://www.worldbank.org/health/impactevaluationtoolkit).

Notas

1. En ausencia de directrices de ética nacional, el investigador y el equipo deberían orientarse según la declaración de Helsinki adoptada por la 29 Asamblea Médica Mundial en Tokio (octubre de 1975) y el artículo 7 del Acuerdo Internacional de Derechos Civiles y Políticos, aprobado por la Asamblea General de las Naciones Unidas el 16 de diciembre de 1966. Se encontrarán otras fuentes en la Organización Mundial de la Salud (OMS) y en el "Informe Belmont sobre principios éticos" y las "Directrices para la protección de seres humanos" (1974) (http://www.hhs.gov/ohrp/policy/belmont.html). Una compilación internacional de normas de investigación sobre seres humanos se puede hallar en http://www.hhs.gov/ohrp/international.
2. Las directrices de la OMS sobre cómo elaborar un protocolo para una investigación que cuente con la participación de seres humanos se pueden encontrar en http://www.who.int/rpc/research_ethics/guide_rp/en/index.html.

3. Para más información sobre los procedimientos de consentimiento durante la recopilación de datos, consúltese el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial.
4. Se puede encontrar más información sobre la asignación de los ID en el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial.
5. Para más información sobre las recomendaciones de la ciencia abierta en el contexto de la evaluación de impacto, consúltese Miguel et al. (2014).
6. Existen otras técnicas. Véase, por ejemplo, Anderson (2008).

Referencias bibliográficas

Anderson, M. L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–95.

Christensen, G. y C. Soderberg. 2015. *The Research Transparency Manual*. Berkeley Initiative for Transparency in the Social Sciences. Disponible en https://github.com/garretchristensen/BestPracticesManual.

Franco, A., N. Malhotra y G. Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343: 30–31.

Vermeersch, C., E. Rothenbühler y J. Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. Washington, D.C.: Banco Mundial. Disponible en http://www.worldbank.org/health/impactevaluationtoolkit.

CAPÍTULO 14

Divulgación de resultados y generación de impacto en las políticas públicas

Una base de evidencia sólida para las políticas públicas

Por fin se ha completado la ardua tarea de evaluar el programa desde el comienzo hasta el final, un esfuerzo de varios años que requirió recursos financieros y humanos considerables. Se han presentado los productos finales de la evaluación, entre ellos un informe de 200 páginas, junto con múltiples anexos. ¿Misión cumplida?

En realidad, ahora se inicia una nueva fase, que consiste en asegurar que todo este esfuerzo rinda sus frutos y se traduzca en un impacto en las políticas. Las evaluaciones de impacto se realizan para rendir cuentas de las inversiones realizadas en el pasado e informar las decisiones de las políticas públicas hacia el futuro pensando en un desarrollo más costo-efectivo, de modo que los escasos recursos produzcan los mayores retornos sociales posibles. Esas decisiones de política pública dependerán de un conjunto de factores, que abarcan desde la economía política hasta las posiciones ideológicas de los usuarios de la información. Sin embargo, las evaluaciones de impacto pueden y deben influir en las políticas públicas proporcionando una sólida base de evidencia que oriente los recursos hacia intervenciones efectivas y probadas.

Desde las primeras etapas de un nuevo programa, incluso cuando este está siendo concebido, la evidencia de evaluaciones de impacto existentes debería desempeñar un rol central para fundamentar el diseño del programa y orientar el próximo conjunto de preguntas de la evaluación.

Concepto clave

Las evaluaciones de impacto deben responder a las preguntas relevantes de las políticas públicas con rigor, aportar evidencia práctica de manera oportuna a los principales interesados, y divulgar la evidencia de forma que sea fácilmente accesible y utilizable por parte de los responsables de las decisiones.

Sin embargo, el proceso de influir en las políticas públicas no suele ocurrir de forma espontánea solo gracias a la generación de evidencia. En primer lugar las evaluaciones de impacto deben responder a preguntas relevantes de las políticas públicas de manera rigurosa, presentando evidencia práctica a los principales interesados de manera oportuna. Sin embargo, puede que los responsables de las políticas y los administradores del programa no tengan ni el tiempo ni la energía para ahondar en los detalles de un informe de 200 páginas, y procuren extraer las principales conclusiones y recomendaciones. La información generada a través de las evaluaciones de impacto tiene que estar organizada y divulgada de manera que sea fácilmente accesible y utilizable para los encargados de la toma de decisiones.

En este capítulo, se trata cómo la evaluación de impacto puede influir en las políticas públicas, los grupos clave a los que conviene llegar, y las estrategias para comunicar y divulgar información para un público focalizado de manera que la evaluación genere un impacto en las políticas públicas.

El punto de partida para influir en las políticas es la selección de las preguntas relevantes de la evaluación que serán útiles para tomar decisiones de políticas públicas, como se señala en la primera parte de este libro. Durante las etapas iniciales del diseño de una evaluación de impacto, es probable que los responsables de las políticas y los evaluadores comiencen con una lista larga de preguntas. Estas preguntas deberían ser consensuadas con el principal grupo de interesados y responsables de las decisiones que, eventualmente, utilizarán la evaluación de impacto para tomar decisiones. La lista larga suele ajustarse y mejorarse con el tiempo para incluir un número más limitado de preguntas bien formuladas que sean relevantes para las políticas públicas y que a su vez puedan responderse mediante una evaluación de impacto, utilizando los métodos expuestos en la segunda parte de este libro. Lograr simultáneamente que los responsables de las políticas definan las preguntas importantes y que el equipo de evaluación pondere la viabilidad técnica de responderlas es un primer paso crucial para influir en las políticas.

Una vez que el programa haya comenzado, es probable que la evaluación de impacto produzca importantes insumos analíticos que pueden servir para fundamentar las políticas mucho antes de que el programa y la evaluación de impacto hayan dado sus frutos. Un ejemplo habitual es el de las conclusiones de una encuesta de línea de base o de un análisis de los resultados a corto plazo. Las encuestas de línea de base a menudo producen los primeros datos exhaustivos y específicos de la población para un programa, y proporcionan estadísticas descriptivas que se pueden incorporar en el diseño del programa y en el

diálogo de políticas. Así como un programa puede tener una descripción general de su población focalizada a través de encuestas nacionales o estudios de diagnóstico, la encuesta de línea de base brinda la primera información detallada sobre subpoblaciones o zonas geográficas específicas donde el programa va a operar. Por ejemplo, un programa diseñado para mejorar la nutrición infantil a través de suplementos nutricionales puede tener estadísticas sobre las tasas de desnutrición crónica y desnutrición aguda a nivel nacional a partir de las encuestas existentes, pero la encuesta de línea de base puede proporcionar las primeras medidas de la situación nutricional y de los hábitos alimentarios del grupo de niños que el programa cubrirá en su área de trabajo. Este tipo de información puede ser valiosa para un diseño de intervención a la medida, y debe hacerse disponible para el equipo de políticas públicas de manera oportuna (idealmente antes de que se implemente la intervención) con el fin de influir en el diseño del programa. El recuadro 14.1 presenta un ejemplo de Mozambique.

Recuadro 14.1: El impacto en las políticas públicas de un modelo innovador de educación preescolar en Mozambique
(continuación del capítulo 1)

En el capítulo 1 (recuadro 1.2) se presentó la evaluación de un programa de educación preescolar comunitario, de Save the Children, aplicado en Mozambique, la cual constituyó un insumo fundamental para las políticas de desarrollo infantil temprano a nivel nacional. Sin embargo, antes de que el programa finalizara, la evaluación generó información nueva y reveladora para el debate de las políticas en este ámbito en el país. La encuesta de línea de base arrojó las primeras medidas de los resultados del desarrollo infantil basadas en la población, utilizando pruebas especializadas de desarrollo infantil adaptadas al contexto de Mozambique, y recopiladas por personal especializado. A pesar de que los datos provenían de un grupo seleccionado de comunidades en una provincia de aquel país, las estadísticas de línea de base proporcionaron una primera imagen de los resultados de desarrollo infantil en el país, mostrando que muchos niños sufrían retrasos en diversos parámetros, desde lenguaje y comunicación hasta desarrollo cognitivo y socioemocional.

La encuesta de línea de base fue presentada por el equipo de evaluación en seminarios y talleres, y se debatieron los resultados con responsables de las políticas de alto nivel, con donantes internacionales y con las principales partes interesadas provenientes de la comunidad de desarrollo infantil temprano (DIT). Los datos generados a través de la evaluación de impacto corroboraron la necesidad de destinar inversiones a este ámbito, y desempeñaron un rol catalítico para movilizar el apoyo a favor de la agenda de desarrollo infantil en el país. Una vez completada, la evaluación se divulgó a través de diversos medios, entre ellos notas informativas de políticas, videos y blogs, algunos de los cuales han sido incorporados en el sitio web de la Iniciativa Internacional para la Evaluación de Impacto (3ie).

Algunas evaluaciones de impacto, sobre todo aquellas que dependen de fuentes de datos administrativos o de encuestas periódicas, pueden producir resultados intermedios que retroalimentan al programa mientras este está siendo implementado. Estos resultados proporcionan información y recomendaciones valiosas sobre cómo los indicadores a lo largo de la trayectoria causal cambian a lo largo del tiempo, lo que permite que se ajusten de manera correspondiente tanto la implementación del programa como la programación en el tiempo de las actividades de evaluación. Por ejemplo, si a mitad de un programa queda claro que no hay efectos en los resultados de corto plazo, puede que se aconseje implementar una evaluación operativa para detectar cuellos de botella y poner en marcha acciones correctivas. El plazo de la evaluación podría ajustarse de modo de evitar la realización de una costosa encuesta de seguimiento antes de que los resultados de la intervención hayan tenido la chance de producirse. En el ejemplo de la nutrición infantil, si los análisis de los datos administrativos sobre la distribución de los suplementos nutricionales demuestran que estos últimos no están llegando a los beneficiarios previstos, el equipo de políticas públicas puede recibir una alerta sobre la necesidad de revisar su cadena de suministro. La encuesta de seguimiento para medir la estatura y el peso de los niños podría aplazarse hasta varios meses después de que el programa haya comenzado a funcionar de manera efectiva, puesto que no hay motivos para creer que el programa nutricional genere impactos antes si no estaba llegando a sus participantes.

Las evaluaciones de impacto tienden a producir grandes volúmenes de información, que abarcan desde los fundamentos técnicos del diseño de evaluación hasta estadísticas descriptivas y análisis de impacto, junto con bases de datos, códigos estadísticos e informes. Es crucial que el equipo de evaluación realice un esfuerzo para documentar toda la información a lo largo del ciclo de evaluación y, en la medida de lo posible, divulgue la documentación (no confidencial) técnica relevante en el dominio público, por ejemplo, a través de un sitio web especializado. Eventualmente, la credibilidad de los resultados de la evaluación dependerá de la metodología y del rigor con que se haya implementado la evaluación. La plena transparencia fortalece la fiabilidad de la evaluación y su potencial para influir en las políticas públicas.

Si bien la completitud y la transparencia son cruciales, la mayoría de los consumidores de la información no ahondarán en detalles. Dependerá del equipo de evaluación elaborar un conjunto manejable de mensajes clave que resuma los resultados y recomendaciones más relevantes para las políticas públicas, y divulgar estos mensajes de forma congruente entre diferentes públicos. La programación de las actividades de divulgación también es esencial para generar un impacto en las políticas. A menos que el equipo de

políticas acuerde lo contrario, las rondas iniciales de presentaciones y consultas sobre los resultados de una evaluación deberían llevarse a cabo internamente, con el personal del programa, los gestores y los responsables de las políticas públicas. Un resultado prematuro filtrado al dominio público puede dañar la reputación de un programa entrañando perjuicios duraderos para el impacto de la evaluación en las políticas.

Elaboración a la medida de una estrategia de comunicación para diferentes públicos

Hay al menos tres públicos primarios para las conclusiones de una evaluación de impacto: el personal del programa y los administradores involucrados en el programa específico que se evalúa; los responsables de las políticas de alto nivel que utilizarán la evaluación para fundamentar las decisiones de financiamiento y de diseño de las políticas; y la comunidad de práctica, que en términos amplios abarca la comunidad académica, los responsables del desarrollo, la sociedad civil (incluidos los medios de comunicación) y los participantes en el programa. Cada uno de estos públicos tendrá diferentes intereses en los resultados de la evaluación y requerirá estrategias de comunicación elaboradas a su medida cuando se trata de conseguir el objetivo de servir de fundamento e influir en las políticas (cuadro 14.1).

Técnicos y administradores. El primer público clave son los miembros del personal técnico y operativo, y los administradores que diseñaron e implementaron el programa, así como los representantes de instituciones (como los ministerios o una institución de financiamiento) estrechamente asociados con el proyecto. Estas personas normalmente serán las primeras en conocer los resultados de la evaluación, y elaborar comentarios sobre las interpretaciones y recomendaciones de la evaluación.

Dado que esta suele ser la primera vez que los resultados ven la luz del día, es clave programar la divulgación de información entre estos interesados. Por un lado, es importante compartir los resultados de forma temprana, de modo que los responsables de las decisiones del programa puedan incorporar cambios y adoptar decisiones de políticas, como aumentar la escala de la intervención (o disminuirla) o ajustar los componentes del programa para mejorar el uso de los recursos y alcanzar un mayor impacto. Por otro lado, hay que hacer una advertencia contra el riesgo de compartir resultados demasiado preliminares basados en un análisis parcial o incompleto, dado que dichos resultados podrían estar sujetos a cambios. Su divulgación podría crear expectativas entre el personal del programa y precipitar decisiones de las políticas aún no maduras que podrían ser caras de revertir en el futuro. Por lo tanto, debería buscarse un equilibrio adecuado de puntualidad y completitud

Cuadro 14.1 Participación de grupos clave en el impacto en las políticas: por qué, cuándo y cómo

	Personal y administradores del programa	Responsables de las políticas de alto nivel	Expertos en desarrollo, académicos, grupos de la sociedad civil
¿Por qué?	Se pueden convertir en defensores de la evaluación de impacto y del uso de evidencia.	Necesitan entender por qué el tema es importante, cómo la evaluación de impacto puede ayudarles a tomar mejores decisiones y, en definitiva, lo que la evidencia les dice acerca de hacia dónde deberían orientarse sus energías (y el financiamiento disponible).	Necesitan evidencia del impacto de los programas de desarrollo con el fin de tomar decisiones, diseñar nuevos programas y llevar a cabo investigación que contribuya a mejorar vidas.
¿Cuándo?	De manera temprana, incluso antes de que se implemente el programa, y con interacciones continuas y frecuentes. Los datos de línea de base se pueden utilizar para elaborar la intervención a la medida. Son los primeros en comentar los resultados de la evaluación.	De manera temprana, al definir las preguntas de la evaluación y antes de que esta comience y, de nuevo, cuando se cuente con los resultados finales. Es importante que los responsables de las políticas públicas comprendan por qué se lleva a cabo una evaluación de impacto y cómo pueden ayudarles los resultados.	De acuerdo con el programa que se evalúe, los grupos de la sociedad civil y los expertos del desarrollo pueden ser importantes defensores locales. La información debería divulgarse una vez que se cuente con los resultados definitivos y estos hayan sido aprobados por el personal del programa y los responsables de las políticas públicas.
¿Cómo?	Introducir el rol de la evidencia en la elaboración de las políticas públicas en un taller para que los administradores del programa participen en el diseño de la evaluación. Efectuar un seguimiento con reuniones en momentos clave: inmediatamente después de la recopilación de datos de línea de base, luego de recopilar resultados intermedios y al final.	Están presentes en los talleres nacionales, y buscan reuniones directas con el personal de nivel superior para explicar el trabajo. Instan a los encargados del programa, al personal técnico y a los responsables de las políticas de nivel medio a mantener a los ministerios informados acerca de la evaluación de impacto. Cuando la evidencia ha acabado, se presenta a los responsables de las políticas de nivel superior. Cuando sea posible, se incluyen análisis de costo-beneficio o costo-efectividad y sugerencias para los próximos pasos.	Los eventos y foros públicos, como seminarios y conferencias, documentos de trabajo, artículos en los periódicos, cobertura en los medios y materiales basados en la red, son medios para llegar a estas audiencias.

en la divulgación inicial de resultados con el equipo del proyecto. Esto suele cumplirse cuando el equipo de evaluación ha llevado a cabo un análisis exhaustivo y verificaciones de robustez, pero antes de que se formulen los resultados, la interpretación y las recomendaciones finales.

Normalmente, al personal y a los encargados del programa les interesarán tanto los detalles técnicos de la metodología de evaluación como también el análisis y los elementos particulares de las conclusiones y recomendaciones presentadas al comienzo. El debate inicial sobre los resultados que se realice con este grupo puede prestarse para reuniones de estilo taller, con presentaciones del equipo de evaluación, y disponiendo de tiempo suficiente para responder a preguntas y comentarios de todas las partes. Este debate inicial suele enriquecer el análisis final, fundamenta la interpretación de resultados y contribuye a elaborar a la medida las recomendaciones finales, de modo que sean más idóneas para orientar los objetivos de las políticas del programa. Los debates iniciales con el personal del programa y los administradores constituyen una buena oportunidad para intercambiar ideas sobre resultados inesperados o potencialmente polémicos, y proponer recomendaciones de políticas públicas y respuestas anticipándose a la divulgación pública de la evaluación de impacto.

Los resultados negativos (incluido el encontrar un impacto nulo) o imprevistos pueden ser decepcionantes para el personal y los administradores del programa que han invertido tiempo y energía considerables, pero también contribuyen a la función crítica de instar a que se reformulen las políticas. Por ejemplo, si se descubre que el programa no ha alcanzado su objetivo primario debido a dificultades en la implementación, se pueden adoptar medidas para abordar esos ámbitos y el programa mejorado se puede volver a evaluar más tarde. Si el programa no produce impactos en el corto plazo o lo hace solamente en un subconjunto de resultados, y hay motivos para creer que se requiere más tiempo para alcanzar los resultados finales, la evaluación puede presentar y defender los resultados iniciales y se pueden planificar otras medidas en una fecha futura. Por último, si está claro que la intervención no consigue generar los beneficios previstos o está provocando un perjuicio inesperado, los administradores del programa pueden tomar medidas inmediatas para detener la intervención o reformular su diseño. De esta manera, cuando se divulgan los resultados de la evaluación, los responsables de las políticas a cargo del programa pueden anunciar medidas correctivas y formular respuestas con antelación, anticipándose a las preguntas difíciles que surgirán en los debates de políticas o en los medios.

Responsables de las políticas de alto nivel. El segundo grupo clave son los responsables de las políticas de alto nivel, que adoptarán decisiones sobre la base de los resultados de las evaluaciones de impacto como, por ejemplo, si ampliar, mantener o disminuir el financiamiento para una intervención.

En este grupo se incluye el Poder Legislativo nacional, los presidentes y primeros ministros, ministros y secretarios principales, juntas de directores y/o donantes. Este grupo de partes interesadas suele contar con los resultados de la evaluación una vez que estos son definitivos y han sido revisados por el personal y los administradores del programa, y aprobados por expertos técnicos externos. En esta etapa, el equipo de evaluación tendrá que centrarse en comunicar los resultados y las recomendaciones clave de manera asequible; los detalles técnicos de la evaluación tienen una importancia secundaria. A los responsables de las políticas de alto nivel les interesará la traducción de los impactos en valores económicamente significativos mediante análisis de costo-beneficio, o una comparación con las intervenciones alternativas a través de análisis de costo-efectividad. Estos parámetros contribuirán a informar a los responsables de las decisiones acerca de si el programa es una manera fiable de invertir recursos limitados para impulsar un objetivo de desarrollo importante. A los responsables de las políticas de alto nivel les puede interesar utilizar los resultados para promover su agenda política, como presionar a favor (o en contra) de una determinada política pública que la evaluación apoya (o no apoya). El equipo de evaluación puede colaborar con los expertos en comunicación para asegurar que los resultados y las recomendaciones relacionadas estén correctamente interpretados y que los mensajes de la estrategia de comunicación sigan alineados con las conclusiones de la evaluación.

La comunidad profesional. El tercer grupo clave para alcanzar un impacto amplio de las políticas son los consumidores de la evaluación fuera del ámbito directo del programa y/o del contexto del país. Este grupo heterogéneo comprende la comunidad profesional en sectores próximos a la evaluación, e incluye a los profesionales del desarrollo, académicos, la sociedad civil y los formuladores de políticas de otros países. Los profesionales del desarrollo más allá del programa específico pueden interesarse en utilizar los resultados de la evaluación para fundamentar el diseño de programas nuevos o existentes. A estos profesionales les interesarán tanto los detalles de la evaluación (métodos, resultados, recomendaciones) como las lecciones operativas y las recomendaciones que puedan contribuir a la implementación de sus propios proyectos de forma más efectiva. Por otro lado, puede que a la comunidad académica le interese más la metodología, los datos y los hallazgos empíricos de la evaluación.

En la sociedad civil destacan dos grupos clave, a saber: los medios y los participantes en el programa. Informar al público de los resultados de una evaluación a través de los medios de comunicación puede desempeñar un rol clave para la rendición de cuentas en materia de gasto público, obtener el apoyo público para las recomendaciones de la evaluación y llevar a cabo políticas efectivas. Esto es particularmente cierto en cuanto a las políticas nuevas

e innovadoras, cuyo resultado era inicialmente incierto u objeto de polémicas en el debate de políticas. Si la evaluación arroja una luz empírica sobre lo que hasta ahora había sido un debate en gran parte teórico o ideológico, puede convertirse en un poderoso instrumento para el cambio de políticas.

Por último, los esfuerzos de divulgación deberían incluir a quienes participan del programa. Los participantes han invertido su tiempo y energía en el programa y puede que hayan dedicado un tiempo considerable a proporcionar información para los fines de la evaluación. Asegurar que tengan acceso a los resultados de la evaluación y que permanezcan informados a propósito de ello es un gesto pequeño pero significativo que puede contribuir a mantener su interés en el programa y a su disposición a tomar parte en futuras evaluaciones.

Divulgación de los resultados

A continuación, se aborda una variedad de estrategias que se pueden considerar para informar a estos grupos clave y generar un impacto en las políticas. Idealmente, las primeras etapas de la planificación de la evaluación incluirán una estrategia de divulgación o de impacto en las políticas. Esta estrategia debería acordarse desde el comienzo, y debería especificar claramente el objetivo de la evaluación para las políticas (por ejemplo, ampliación de un modelo de intervención más costo-efectivo), el público clave al que la evaluación intenta llegar, las estrategias de comunicación usadas y un presupuesto para realizar actividades de divulgación. Si bien el formato y contenido de las actividades y de los productos de la divulgación variarán según cada caso, en el resto de este capítulo se presentan algunas sugerencias y orientaciones generales. El recuadro 14.2 incluye una lista de algunos instrumentos de extensión y divulgación.

Los informes suelen ser el primer medio para divulgar el conjunto completo de resultados de la evaluación. Se recomienda que estos informes tengan una extensión moderada, entre 30 y 50 páginas, e incluyan un resumen de una página, o menos, y un resumen ejecutivo de dos a cuatro páginas con los principales resultados y recomendaciones. Los detalles técnicos, la documentación relacionada y el análisis de apoyo como pruebas de robustez y falsificación se pueden presentar en anexos o apéndices.

La publicación de una evaluación de impacto como documento de trabajo académico y/o artículo en una revista científica arbitrada puede ser una medida final laboriosa pero muy provechosa para presentar los resultados de la evaluación. Las rigurosas revisiones de pares requeridas para el proceso de publicación proporcionarán una retroalimentación valiosa que mejorará el análisis y la interpretación de los resultados, y la publicación

Recuadro 14.2: Instrumentos de extensión y divulgación

A continuación se listan algunos ejemplos de medios para divulgar las evaluaciones de impacto:

- Exposiciones sobre el programa y resultados de la evaluación.
- Videos donde los beneficiarios dan su opinión del programa y revelan cómo afecta sus vidas.
- Breves notas informativas en las cuales se explica la evaluación y se resumen las recomendaciones de políticas.
- Blogs de los investigadores y responsables de las políticas que explican la importancia de la evaluación.
- Informes completos, después de recibir los resultados finales, con exhaustivos resúmenes ejecutivos para asegurar que los lectores entiendan rápidamente las principales conclusiones.
- Invitaciones para los medios que permitan a los periodistas ver el programa en acción y los resultados del informe.

puede transmitir una clara señal a los responsables de las políticas sobre la calidad y credibilidad de los resultados de una evaluación.

Sobre la base de la estrategia de divulgación acordada, los informes y documentos se pueden publicar en diversos medios, entre ellos, el sitio web del programa, el sitio web de la institución evaluadora, como parte de una serie de documentos de trabajo, revistas académicas arbitradas y libros.

Si bien los informes de evaluación y los documentos académicos sirven como fundamento para la estrategia de divulgación, su alcance entre un público más amplio fuera de la comunidad profesional y académica puede ser limitado debido a su extensión y a su lenguaje técnico. Puede que el equipo de evaluación, quizás en colaboración con los expertos en comunicación, considere útil producir artículos breves, escritos al estilo de un relato o con un estilo periodístico, con un lenguaje claro y sencillo para llegar a públicos más amplios. Se pueden dar a conocer artículos breves bajo la forma de notas informativas de políticas, boletines e infografías. En estas publicaciones, será particularmente útil eliminar la jerga técnica y traducir los resultados en representaciones visualmente atractivas, con imágenes, esquemas y gráficos (recuadro 14.3).

Los equipos de evaluación pueden generar un conjunto de presentaciones que acompañen a los informes escritos y los artículos breves. Las presentaciones deberían elaborarse a la medida del público específico. Un buen punto de partida es producir una presentación técnica para el equipo del proyecto y el público académico, y otra presentación más breve y menos

Recuadro 14.3: La divulgación efectiva de las evaluaciones de impacto

Diversas publicaciones exponen los resultados de las evaluaciones de impacto en un formato accesible y sencillo. Entre ellos se incluyen dos actualizaciones con un foco regional.

- Los resultados de la evaluación de impacto de programas en América Latina y el Caribe se recogen en el *Panorama de la efectividad en el desarrollo* (DEO, por sus siglas en inglés), publicado anualmente por la Oficina de Planificación Estratégica y Efectividad en el Desarrollo del Banco Interamericano de Desarrollo (BID). Los resultados se resumen en artículos breves, de fácil lectura, que incluyen resúmenes infográficos de una página que explican la pregunta fundamental de la evaluación de impacto, los métodos, resultados y recomendaciones de políticas, utilizando gráficos e íconos que permiten que los lectores entiendan los mensajes clave de forma muy rápida e intuitiva. El DEO de 2014 incluye los resultados de evaluaciones de impacto de programas tan diversos como el turismo en Argentina, la capacitación laboral en República Dominicana, la productividad agrícola en Bolivia y las orquestas juveniles en Perú.
- *Africa Impact Evaluation Update*, del Banco Mundial, recoge la última evidencia de la región. En 2013, esta publicación se centró en el género y en 2014, en la agricultura y la tierra.

Fuentes: http://deo.iadb.org y http://www.worldbank.org.

técnica para los responsables de las políticas y la sociedad civil. Si bien las principales conclusiones y recomendaciones para las políticas serán las mismas, la estructura y el contenido de estas dos presentaciones tendrán diferencias importantes. La presentación técnica debería centrarse en afianzar la credibilidad de los resultados mediante una exposición de los métodos de evaluación, los datos y el análisis, antes de llegar a los resultados y recomendaciones. Una presentación dirigida a los responsables de las políticas debería poner de relieve el problema del desarrollo que la intervención se propone abordar y las implicaciones prácticas de las conclusiones, y tratar de forma más superficial los detalles técnicos.

Para aprovechar el cada vez mayor acceso a Internet de los países en desarrollo y las alternativas de bajo costo para producir multimedia, los equipos de evaluación también pueden contemplar una gama de medios para divulgar las conclusiones de la evaluación, entre ellos: los sitios web o las grabaciones en audio y video. Los videoclips de corta duración pueden ser un medio poderoso para transmitir ideas complejas a través de imágenes y sonido, dejando que la historia de la evaluación se despliegue de una manera que sea más rápida y más plenamente comprensible que la que utilizan los típicos medios impresos (recuadro 14.4).

Por último, armado con una variedad de productos de divulgación, el equipo de evaluación debe mostrarse proactivo en la divulgación de estos productos a los consumidores dentro del programa, del gobierno y de la comunidad profesional más amplia, de modo que la información llegue a los usuarios previstos y pueda ser asimilada en el proceso de toma de decisiones y el debate de políticas públicas. El proceso de divulgación se lleva a cabo mediante reuniones presenciales entre el equipo de evaluación y el administrador del programa, a través del cabildeo con responsables de las políticas de alto nivel, así como también de presentaciones en seminarios y conferencias donde los académicos y miembros de la comunidad profesional se reúnen para informarse acerca de los últimos avances en la investigación y la evaluación del desarrollo, mediante entrevistas y programas de noticias en la radio y la televisión y, actualmente cada vez más, a través de Internet. Los blogs y las redes sociales en particular pueden ser maneras costo-efectivas de llegar a grandes cantidades de usuarios potenciales y para orientar a los lectores hacia un conjunto de productos disponibles relacionados con una determinada evaluación (recuadro 14.5). Si bien las estrategias particulares variarán según cada caso, se recomienda una vez más planificar y presupuestar los medios y las actividades de divulgación con antelación, de modo que los resultados de la evaluación puedan llegar a sus públicos previstos de manera rápida y efectiva, de modo que así se pueda maximizar el impacto en las políticas.

Recuadro 14.4: Divulgación de las evaluaciones de impacto en línea

A continuación, se muestran algunos ejemplos destacados de divulgación en línea de los resultados de una evaluación de impacto:

- La Iniciativa Internacional para la Evaluación de Impacto (3ie) organiza la evidencia de las evaluaciones de impacto por sector, e incluye notas informativas de política, revisiones sistemáticas y mapas de brechas de evidencia.
- El Abdul Latif Jameel Poverty Action Lab (J-Pal) divulga evidencia de evaluaciones de impacto realizadas por investigadores asociados, y añade notas informativas de políticas, análisis de costo-efectividad y enlaces con documentos académicos.
- La Iniciativa de Desarrollo de la Evaluación de Impacto (DIME, por sus siglas en inglés) del Banco Mundial presenta notas breves, boletines e informes con los resultados de las evaluaciones de impacto de los proyectos de dicha institución.
- El Fondo Estratégico para la Evaluación de Impacto (SIEF) del Banco Mundial incluye videos, notas breves y entrevistas.

Recuadro 14.5: Blogs de evaluación de impacto

Esta lista contiene ejemplos de blogs que recogen con regularidad los resultados de las evaluaciones de impacto:

- El blog de Impacto en el Desarrollo del Banco Mundial.
- El blog de Efectividad en el Desarrollo del BID.
- El blog de Innovations for Poverty Action.

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- La Iniciativa Internacional para la Evaluación de Impacto (3ie) y el Instituto de Desarrollo de Ultramar (ODI, por sus siglas en inglés) han desarrollado un kit de herramientas de impacto de políticas en línea con el fin de contribuir a divulgar y utilizar la evidencia de las evaluaciones de impacto en la toma de decisiones.

Cuarta parte

CÓMO OBTENER DATOS PARA UNA EVALUACIÓN DE IMPACTO

La cuarta parte de este libro proporciona orientación sobre cómo obtener datos para una evaluación de impacto, lo que comprende la elección de la muestra y cómo encontrar fuentes de datos adecuadas.

El capítulo 15 trata de cómo extraer una muestra de una población de interés y cómo llevar a cabo cálculos de potencia para determinar el tamaño adecuado de la muestra de la evaluación de impacto. El capítulo se centra en la descripción de la idea fundamental de los muestreos y los cálculos de potencia.

También destaca los elementos que los responsables de las políticas tienen que proporcionar al equipo de investigación o a los expertos técnicos responsables de elaborar los muestreos y los cálculos de potencia.

En el capítulo 16 se analizan las diversas fuentes de datos que pueden utilizar las evaluaciones de impacto. Allí se destaca cuándo se pueden usar las fuentes de los datos existentes, entre ellos los datos administrativos. Dado que numerosas evaluaciones requieren la recopilación de nuevos datos, en el capítulo se abordan los pasos necesarios para recopilar los datos de una nueva encuesta. Esto implica determinar quién recopilará los datos, desarrollar instrumentos de recopilación de datos y realizar pruebas piloto, llevar a cabo el trabajo de campo y de control de calidad, y procesar y almacenar datos.

En el capítulo 17 se presentan conclusiones de la totalidad del libro. Allí se revisan brevemente los elementos centrales de una evaluación de impacto bien diseñada, y se proponen algunos consejos para mitigar los riesgos habituales en la realización de una evaluación de impacto. También se ofrecen algunas perspectivas del reciente aumento del uso de evaluaciones de impacto y otras iniciativas de institucionalización relacionadas.

CAPÍTULO 15

La elección de una muestra

El muestreo y los cálculos de potencia

Una vez que se ha elegido el método para seleccionar el grupo de comparación y estimar el contrafactual, uno de los próximos pasos consiste en determinar qué datos se precisarán, y la muestra necesaria para estimar con exactitud las diferencias de los resultados entre el grupo de tratamiento y el grupo de comparación. En este capítulo, se analizará cómo se puede extraer una muestra de una población de interés (muestreo) y cómo se puede determinar el tamaño que debe tener la muestra para proporcionar estimaciones precisas del impacto del programa (cálculos de potencia). El muestreo y los cálculos de potencia requieren habilidades técnicas específicas y se les suelen encargar a un experto especializado. En este capítulo, se describen los elementos básicos de la realización de muestreos y cálculos de potencia, y se destacan los elementos que los responsables de las políticas deben poder proveer a los expertos técnicos.

Elaboración de una muestra

El *muestreo* es el proceso de extraer unidades de una población de interés para estimar las características de la población. Suele ser necesario, dado que, normalmente, no es posible observar y medir directamente los resultados para toda la población de interés. Por ejemplo, si se desea conocer la altura

promedio de los niños menores de 2 años en un país, sería muy difícil, costoso y lento medir a todos los niños de la población. En cambio, se puede utilizar una muestra de niños extraída de la población para inferir las características promedio de esa población (gráfico 15.1).

El proceso mediante el cual se extrae una muestra de la población de interés es crucial. Los principios de muestreo sirven de orientación para extraer muestras representativas. En la práctica, hay que seguir tres grandes pasos para extraer una muestra:

1. Determinar la población de interés.
2. Definir un marco muestral.
3. Extraer el número de unidades requeridas por los cálculos de potencia del marco muestral.

En primer lugar, se debe definir claramente la *población de interés*. Esto requiere especificar con precisión la unidad en la población de interés para la cual se medirán los resultados, y detallar con claridad la cobertura geográfica o cualquier otro atributo pertinente que caracterice a la población de interés. Por ejemplo, si se está gestionando un programa de desarrollo infantil temprano, puede que resulte de interés medir el impacto del programa en los resultados cognitivos de los niños de entre 3 y 6 años en todo el país, solo para los niños que viven en zonas rurales o solo para los niños matriculados en preescolar.

Gráfico 15.1 Uso de una muestra para inferir las características promedio de una población de interés

En segundo lugar, una vez que se haya definido la población de interés, se debe establecer un *marco muestral*. El marco muestral es la lista más exhaustiva que se puede obtener de las unidades en la población de interés. Idealmente, el marco muestral debería coincidir exactamente con la población de interés. Un censo totalmente actualizado de la población de interés constituiría un marco muestral ideal. En la práctica, se suelen utilizar como marcos muestrales las listas existentes, como los censos de población, los censos de instalaciones o los registros de inscritos.

Se requiere un marco muestral adecuado para asegurar que las conclusiones a las que se llegue mediante el análisis de una muestra se puedan generalizar para el conjunto de la población. De hecho, un marco muestral que no coincida exactamente con la población de interés crea un *sesgo de cobertura*, como lo ilustra el gráfico 15.2. Si se produce un sesgo de cobertura, los resultados de la muestra no tienen validez externa para el conjunto de la población de interés sino únicamente para la población incluida en el marco muestral. La medida en que las estadísticas calculadas a partir de la muestra se pueden generalizar a toda la población de interés depende de la magnitud del sesgo de cobertura, es decir: de la falta de coincidencia entre el marco muestral y la población de interés.

Concepto clave

Un marco muestral es la lista más exhaustiva que se puede obtener de las unidades en la población de interés. Se produce un ***sesgo de cobertura*** cuando el marco muestral no corresponde perfectamente a la población de interés.

Los sesgos de cobertura constituyen un riesgo, y la creación de marcos muestrales requiere un esfuerzo riguroso. Por ejemplo, los datos del censo pueden contener la lista de todas las unidades de una población. Sin embargo, si ha transcurrido demasiado tiempo entre el censo y el momento

Gráfico 15.2 Un marco muestral válido cubre el conjunto de la población de interés

en que se recopilaron los datos de la muestra, el marco muestral ya no estará actualizado. Además, los datos del censo pueden no contener suficiente información sobre atributos específicos para construir un marco muestral. Si la población de interés está compuesta por niños que asisten a nivel inicial, y el censo no incluye datos sobre la matrícula preescolar, se requerirían datos complementarios de matrícula o registros de los establecimientos educativos.

Concepto clave

El muestreo es el proceso por el cual las unidades se extraen de un marco muestral. El muestreo probabilístico asigna una probabilidad bien definida a cada unidad del marco muestral.

Una vez identificada la población de interés y un marco muestral, es necesario elegir un método para elaborar la muestra. Se pueden utilizar diversos procedimientos alternativos.

Los métodos de *muestreo probabilístico* son los más rigurosos, dado que asignan una probabilidad bien definida para cada unidad del marco muestral. Los tres principales métodos de muestreo probabilístico son los siguientes:

- *Muestreo aleatorio.* Todas las unidades de la población tienen exactamente la misma probabilidad de ser extraídas.[1]

- *Muestreo aleatorio estratificado.* La población se divide en dos grupos (por ejemplo, hombres y mujeres) y se lleva a cabo un muestreo aleatorio en cada grupo. Como consecuencia, todas las unidades en cada grupo (o estrato) tienen la misma probabilidad de ser extraídas. Siempre y cuando todos los grupos sean lo suficientemente grandes, el muestreo estratificado permite formular inferencias acerca de los resultados no solo a nivel de la población, sino también dentro de cada grupo. El muestreo estratificado es útil cuando se quiere elaborar una muestra de los subgrupos pequeños en la población (por ejemplo, las minorías) con el fin de estudiarlos más en detalle. La estratificación es esencial para las evaluaciones que buscan comparar los impactos del programa entre esos subgrupos.

- *Muestreo de* clusters. Las unidades se agrupan en *clusters* (conglomerados) y se extrae una muestra aleatoria de los mismos. Posteriormente, o todas las unidades en esos *clusters* constituyen la muestra, o bien se extrae un cierto número de unidades del *cluster* de forma aleatoria. Esto significa que cada *cluster* tiene una probabilidad bien definida de ser seleccionado y las unidades dentro de un *cluster* seleccionado también tienen una probabilidad bien definida de ser extraídas.

En el contexto de una evaluación de impacto, el procedimiento para extraer una muestra a menudo está determinado por las reglas de elegibilidad del programa que se evalúa. Como se verá en el debate sobre el tamaño de la muestra, si la unidad viable más pequeña de implementación es más grande que la unidad de observación, la asignación aleatoria de los

beneficios creará *clusters*. Por este motivo, el muestreo de *clusters* aparece a menudo en los estudios de evaluaciones de impacto.

El *muestreo no probabilístico* puede provocar graves errores de muestreo. Por ejemplo, supóngase que se emprende una encuesta nacional pidiendo a un grupo de entrevistadores que recopilen datos de los hogares de las viviendas más próximas a la escuela en cada pueblo. Cuando se utiliza un procedimiento de muestreo no probabilístico de este tipo, es probable que la muestra no sea representativa del conjunto de la población de interés. Concretamente, se producirá un sesgo de cobertura, dado que las viviendas remotas no serán estudiadas.

Es necesario prestar mucha atención al marco muestral y al procedimiento de muestreo para establecer si los resultados obtenidos de una determinada muestra se pueden generalizar al conjunto de la población de interés. Aun cuando el marco muestral tenga perfecta cobertura y se utilice un procedimiento de muestreo probabilístico, los errores de no muestreo también pueden afectar la validez interna y externa de la evaluación de impacto. Los errores de no muestreo se tratan en el capítulo 16. Por último, en ocasiones se observa una confusión entre el muestreo aleatorio y la asignación aleatoria. En el recuadro 15.1 se explica con claridad que ambos son muy diferentes.

En el resto de este capítulo, se examina la importancia que entraña el tamaño de la muestra para la precisión de las evaluaciones de impacto. Como se verá con mayor exactitud, se requieren muestras relativamente

Recuadro 15.1: El muestreo aleatorio no es suficiente para la evaluación de impacto

A veces se produce una confusión entre el muestreo aleatorio y la asignación aleatoria. ¿Qué pasaría si alguien comenta que está implementando una evaluación de impacto entrevistando a una *muestra aleatoria* de participantes y no participantes? Supóngase que observa a un grupo de individuos que participan de un programa de empleo y a un grupo de individuos que no participan en el programa. ¿Qué pasaría si se tomara una muestra aleatoria de cada uno de estos dos grupos? El primer gráfico ilustra que se obtendría una muestra aleatoria de participantes y una muestra aleatoria de no participantes. Si los participantes y los no participantes tienen diferentes características, también lo tendrá la muestra de participantes y no participantes. El muestreo aleatorio no hace que dos grupos no comparables sean comparables y no proporciona validez interna para la evaluación de impacto. Este es el motivo por el que el muestreo aleatorio no es suficiente para la evaluación de impacto.

Continúa en la página siguiente.

Recuadro 15.1: El muestreo aleatorio no es suficiente para la evaluación de impacto *(continúa)*

Gráfico B15.1.1 Muestreo aleatorio entre grupos no comparables de participantes y no participantes

Como debería quedar claro a partir del debate que se desarrolla en la segunda parte, la asignación aleatoria de los beneficios de un programa es diferente del muestreo aleatorio. El proceso de asignación aleatoria parte de una población de interés elegible y utiliza un procedimiento de aleatorización para asignar las unidades (que normalmente son personas o grupos de personas, como niños en una escuela) de la población elegible a un grupo de tratamiento que será objeto de una intervención, y a un grupo de comparación que no lo será. El proceso de aleatorización de un programa que se exhibe en el gráfico B15.1.2 es diferente del proceso de muestreo aleatorio descrito en el gráfico B15.1.1. Como se señaló en la segunda parte, cuando la asignación aleatoria está bien implementada, contribuye a la validez interna de la evaluación de impacto. El muestreo aleatorio puede ser útil para asegurar la validez externa, en la medida en que la muestra se extrae aleatoriamente de la población de interés.

Continúa en la página siguiente.

Recuadro 15.1: El muestreo aleatorio no es suficiente para la evaluación de impacto *(continúa)*

Gráfico B15.1.2 Asignación aleatoria de los beneficios de un programa entre un grupo de tratamiento y un grupo de comparación

más grandes para obtener estimaciones precisas de las características de la población. También se requieren muestras más grandes para poder obtener estimaciones precisas de las diferencias entre grupos de tratamiento y de comparación, es decir, para estimar el impacto de un programa.

La decisión sobre el tamaño de la muestra de una evaluación de impacto: cálculos de potencia

Como ya se señaló, el muestreo describe el proceso para elaborar una muestra de unidades de una población de interés a fin de estimar las características de esa población. Las muestras más grandes dan estimaciones más precisas de las características de la población. ¿De qué tamaño, exactamente, tienen que ser las muestras para una evaluación de impacto?

Los cálculos para determinar el tamaño de la muestra se denominan cálculos de potencia. Aquí se analiza la idea básica que subyace a los cálculos de potencia a partir del caso más sencillo, a saber: una evaluación realizada utilizando un método de asignación aleatoria, para probar la efectividad de un programa en relación con un grupo de comparación que no recibe una intervención, y suponiendo que el incumplimiento no es un problema.[2] Al final del capítulo, se abordan brevemente otras consideraciones más allá de este caso sencillo.

El fundamento de los cálculos de potencia

Los *cálculos de potencia* indican el tamaño mínimo de la muestra que es necesario para llevar a cabo una evaluación de impacto y para responder de forma convincente a la pregunta de interés para las políticas. Concretamente, los cálculos de potencia se pueden utilizar para:

- Evaluar si las bases de datos existentes son suficientemente grandes para llevar a cabo una evaluación de impacto.
- Evitar recopilar pocos datos. Si la muestra es demasiado pequeña, puede que no sea posible detectar un impacto positivo –aunque existiera– y, por lo tanto, se puede llegar a la conclusión de que no ha tenido efecto. Esto podría provocar una decisión de política para eliminar el programa, lo cual sería perjudicial.
- Contribuir a tomar decisiones a propósito del tamaño adecuado de la muestra. Los tamaños más grandes de la muestra proporcionan estimaciones más precisas de los impactos del programa, pero la recopilación de información puede ser muy onerosa. Los cálculos de potencia proporcionan insumos clave para evaluar el equilibrio entre los costos requeridos para recopilar más datos y los beneficios de una mayor precisión en la evaluación de impacto.

Concepto clave
Los cálculos de potencia proporcionan un indicador de la muestra más pequeña con la que es posible estimar con precisión el impacto de un programa; a saber, la muestra más pequeña que permitirá detectar diferencias significativas en los resultados entre los grupos de tratamiento y comparación.

Los cálculos de potencia constituyen una indicación de la muestra más pequeña (y el presupuesto más bajo) con el que es posible medir el impacto de un programa; es decir, la muestra más pequeña que permitirá detectar diferencias significativas en los resultados entre los grupos de tratamiento y comparación. Por lo tanto, los cálculos de potencia son cruciales para determinar cuáles son los programas que tienen éxito y cuáles no.

Como se señaló en el capítulo 1, la pregunta básica de la evaluación de impacto es: ¿Cuál es el impacto o efecto causal de un programa en un resultado de interés? La sencilla hipótesis incorporada en esa pregunta puede ser reformulada de la siguiente manera: ¿El impacto del programa es diferente

de cero? En el caso de la asignación aleatoria, responder a esta pregunta requiere dos pasos:

1. Estimar los resultados promedio para los grupos de tratamiento y comparación.
2. Valorar si existe una diferencia entre el resultado promedio del grupo de tratamiento y el resultado promedio del grupo de comparación.

A continuación, se analizará cómo calcular los resultados promedio para cada grupo, y luego, cómo comprobar si hay una diferencia entre los dos grupos.

Estimación de resultados promedio para los grupos de tratamiento y comparación

Supóngase que se debe estimar el impacto de un programa de nutrición en el peso de los niños a los 2 años, y que hay 200.000 niños elegibles para el programa. Del total de niños elegibles, 100.000 fueron asignados de forma aleatoria para participar en el programa. Los 100.000 niños elegibles que no fueron asignados aleatoriamente al programa sirven como grupo de comparación. Como primer paso, habrá que estimar el peso promedio de los niños que participaron y de los que no participaron.

Para determinar el peso promedio de los niños que participaron, se podría pesar a cada uno de los 100.000 niños participantes y luego calcular el promedio. Desde luego, sería un procedimiento sumamente costoso. Afortunadamente, no es necesario pesar a cada niño. El promedio se puede estimar utilizando el peso promedio de una muestra extraída de la población de los niños que participan.[3] Cuantos más niños haya en la muestra, más cerca estará el promedio estimado del promedio real. Cuando una muestra es pequeña, el peso promedio constituye una estimación muy imprecisa del promedio en la población. Por ejemplo, una muestra de dos niños no dará una estimación precisa. En cambio, una muestra de 10.000 niños producirá una estimación más precisa mucho más cercana al verdadero peso promedio. En general, cuantas más observaciones haya en la muestra, más precisas serán las estadísticas obtenidas de la muestra (gráfico 15.3).[4]

Por lo tanto, se sabe que con una muestra más grande se obtendrá una imagen más exacta de la población de los niños que participan. Lo mismo ocurrirá con los niños que no participan: a medida que crece el tamaño de la muestra de estos últimos, se sabe con mayor precisión cómo es esa población. ¿Pero por qué habría esto de importar? Si se puede estimar el resultado promedio (el peso) de los niños que participan y no participan

Gráfico 15.3 Una muestra más grande tiene más probabilidades de parecerse a la población de interés

con más precisión, también se podrá saber con más precisión la diferencia de peso entre ambos grupos, y eso es el impacto del programa. Dicho de otra manera, si solo se tiene una idea vaga del peso promedio de los niños en los grupos de pequeños que participan (tratamiento) y que no participan (comparación), ¿cómo se podrá tener una idea precisa de la diferencia de peso de los dos grupos? La verdad es que no se puede. En la siguiente sección, se examina esta idea de una manera ligeramente más formal.

Comparación de los resultados promedio entre los grupos de tratamiento y comparación

Una vez que se haya estimado el resultado promedio (el peso) del grupo de tratamiento (los niños que participan seleccionados por asignación aleatoria) y el grupo de comparación (los niños que no participan seleccionados por asignación aleatoria), se puede proceder a determinar si los dos resultados son diferentes. Esta parte está clara: se restan los promedios y se calcula la diferencia. En términos estadísticos, la evaluación de impacto pone a prueba la *hipótesis nula* (o *por defecto*) en contraste con la *hipótesis alternativa*.

La hipótesis nula es la hipótesis de que el programa no tiene un impacto. Se expresa como:

H_0: impacto o diferencia entre el resultado en el grupo de tratamiento y comparación = 0.

H_a: impacto o diferencia entre el resultado en el grupo de tratamiento y comparación ≠ 0.

Imagínese que en un ejemplo de un programa de nutrición se comienza con una muestra de dos niños tratados y dos niños de comparación. Con una muestra tan pequeña, la estimación del peso promedio de los niños tratados y los niños de comparación y, por lo tanto, la estimación de la diferencia entre los dos grupos, no será demasiado fiable. Puede verificarse esto extrayendo diferentes muestras de dos niños del grupo de tratamiento y dos niños del grupo de comparación. Lo que se encontrará es que el impacto estimado del programa varía mucho.

Al contrario, imagínese que se comienza con una muestra de 1.000 niños tratados y 1.000 niños del grupo de comparación. Como se señaló, las estimaciones del peso promedio de ambos grupos serán mucho más precisas. Por lo tanto, la estimación de la diferencia entre los dos grupos también lo será.

Por ejemplo, figúrese que se observa que el peso promedio en la muestra de los niños del tratamiento (que participan) es de 12,2 kilos, y el promedio de los niños en la muestra de comparación (que no participan) es de 12,0 kilos. La diferencia entre ambos grupos es de 0,2 kilos. Si estas cifras correspondieran a muestras de dos observaciones cada una, no se sabría bien si el impacto del programa es verdaderamente positivo porque esos 0,2 kilos podrían deberse a la falta de precisión en las estimaciones. Sin embargo, si estas cifras provienen de muestras de 1.000 observaciones cada una, aumentaría la confianza de que se acercan bastante al verdadero impacto del programa, que en este caso sería positivo.

Por lo tanto, la pregunta clave es: ¿Exactamente qué tamaño debe tener la muestra para permitirnos saber que un impacto estimado positivo se debe al verdadero impacto del programa y no a una falta de precisión en las estimaciones?

Dos errores potenciales en las evaluaciones de impacto

Cuando se prueba si un programa tiene impacto, se pueden cometer dos tipos de errores. Se comete un *error de tipo I* si una evaluación concluye que el programa ha tenido impacto, cuando en realidad no lo ha tenido. En el caso de la intervención hipotética en nutrición, esto ocurriría si usted, como miembro del equipo de evaluación, concluyera que el peso promedio de los

Concepto clave

Un *error de tipo I* ocurre cuando una evaluación llega a la conclusión de que un programa ha tenido impacto, cuando en realidad no lo ha tenido. Un *error de tipo II* se produce cuando una evaluación llegará la conclusión de que el programa no ha tenido impacto cuando, de hecho, sí lo ha tenido.

niños de la muestra tratada es superior al de los niños de la muestra de comparación, aunque el peso promedio de los pequeños en las dos poblaciones es, de hecho, igual y las diferencias observadas eran pura coincidencia. En este caso, el impacto positivo que se observó provendría únicamente de la falta de precisión de las estimaciones.

Un *error de tipo II* es el tipo contrario de error. Se produce cuando una evaluación llega a la conclusión de que el programa no ha tenido impacto, cuando en realidad sí lo ha tenido. En el caso de la intervención en nutrición, esto ocurriría si se concluyera que el peso promedio de los niños en las dos muestras es el mismo, aunque el peso promedio de los niños de la población de tratamiento es, de hecho, superior al de los niños del grupo de comparación. Una vez más, el impacto debería haber sido positivo, pero debido a la falta de precisión de las estimaciones, se llega a la conclusión de que el programa ha tenido un impacto cero.

Cuando se prueba la hipótesis de que un programa ha tenido impacto, los estadísticos pueden limitar el tamaño de los errores de tipo I. La probabilidad de un error de tipo I se puede establecer mediante un parámetro denominado el *nivel de significancia*. El nivel de significancia suele fijarse en 5%, lo que quiere decir que se puede tener un 95% de confianza en llegar a la conclusión de que el programa ha tenido un impacto. Si a usted le preocupa mucho cometer un error de tipo I, puede establecer un nivel de significancia menor: por ejemplo, del 1%, de manera de tener un 99% de confianza de llegar a la conclusión de que el programa ha tenido impacto.

Sin embargo, los errores de tipo II también preocupan a los responsables de las políticas. Numerosos factores influyen en la probabilidad de cometer un error de tipo II, pero el tamaño de la muestra es crucial. Si el peso promedio de 50.000 niños tratados es el mismo que el peso promedio de 50.000 niños de comparación, es probable que se pueda concluir que el programa no ha tenido impacto. Al contrario, si en una muestra de dos niños del grupo de tratamiento estos pesan en promedio lo mismo que en el caso de la muestra de dos niños del grupo de comparación, es más difícil llegar a una conclusión fiable. ¿El peso promedio es similar porque la intervención ha tenido impacto o porque los datos no son suficientes para comprobar la hipótesis en una muestra tan pequeña? Las muestras grandes reducen la probabilidad de que solo se observe a los niños que pesan lo mismo por una cuestión de (mala) suerte. En las muestras grandes, la diferencia de promedios entre la muestra tratada y la muestra de comparación proporciona una mejor estimación de la verdadera diferencia de los promedios entre todas las unidades tratadas y todas las unidades de comparación.

Concepto clave

La potencia es la probabilidad de detectar un impacto cuando, de hecho, este existe. Una evaluación de impacto tiene una alta potencia si hay un bajo riesgo de que no se detecten los impactos reales del programa; es decir, de cometer un error de tipo II.

La *potencia* (o *potencia estadística*) de una evaluación de impacto es la probabilidad de detectar una diferencia entre los grupos de tratamiento y comparación cuando esta de hecho existe. Una evaluación de impacto tiene

una alta potencia si hay un bajo riesgo de no detectar verdaderos impactos del programa, es decir, de cometer un error de tipo II. Los ejemplos anteriores muestran que el tamaño de la muestra es un factor determinante crucial de la potencia de una evaluación de impacto. Las secciones siguientes ilustrarán más detenidamente este punto.

Por qué los cálculos de potencia importan en las políticas públicas

El objetivo del cálculo de potencia consiste en determinar el tamaño de una muestra para evitar llegar a la conclusión de que un programa no ha tenido impacto, cuando de hecho sí lo ha tenido (error de tipo II). La potencia de una prueba es igual a 1 menos la probabilidad de un error de tipo II.

Una evaluación de impacto tiene una *potencia elevada* si es poco probable que se produzca un error de tipo II, lo que significa que es poco probable que usted se sienta decepcionado por los resultados que muestran que el programa que se evalúa no ha tenido impacto, cuando en realidad sí lo ha tenido.

Desde una perspectiva de políticas, las *evaluaciones de impacto con insuficiente potencia,* con una alta probabilidad de errores de tipo II, no solo son inútiles sino que también pueden resultar muy onerosas. Una alta probabilidad de un error de tipo II pone en peligro el potencial de una evaluación de impacto de identificar resultados estadísticamente significativos. Por lo tanto, destinar recursos a evaluaciones de impacto sin suficiente potencia es una inversión riesgosa.

Las evaluaciones de impacto sin suficiente potencia también pueden tener graves consecuencias prácticas. Por ejemplo, en la intervención hipotética en nutrición anteriormente mencionada, si se llegara a la conclusión de que el programa no fue efectivo, aunque sí lo fue, los responsables de las políticas podrían poner fin a un programa que, de hecho, beneficia a los niños. Por lo tanto, es crucial minimizar la probabilidad de errores de tipo II utilizando muestras lo suficientemente grandes en las evaluaciones de impacto. Por esto es tan fundamental y pertinente llevar a cabo cálculos de potencia.

Los cálculos de potencia paso a paso

A continuación, se explican los principios básicos de los cálculos de potencia, con eje en el caso sencillo de un programa de asignación aleatoria. Para llevar a cabo cálculos de potencia se requiere estudiar las siguientes cinco preguntas:

1. ¿El programa funciona mediante *clusters*?
2. ¿Cuál(es) es/son los *indicadores de resultados*?

3. ¿Cuál es el *nivel mínimo de impacto* que justificaría la inversión hecha en la intervención?

4. ¿Cuál es la media de resultado para la población de interés? ¿Cuál es la *varianza subyacente* del indicador de resultado?

5. ¿Cuáles son los niveles razonables de *potencia estadística* y de *significancia estadística* en la evaluación que se lleva a cabo?

Cada una de estas preguntas es válida para el contexto específico de las políticas en el que se ha decidido llevar a cabo la evaluación de impacto.

El primer paso en los cálculos de potencia consiste en determinar si el programa que se quiere evaluar genera *clusters* a lo largo de su implementación. Una intervención cuyo nivel de intervención (a menudo, lugares) es diferente del nivel al que se querrían medir los resultados (a menudo, personas) genera *clusters* en torno al lugar de la intervención. Por ejemplo, puede que sea necesario implementar un programa en el nivel del hospital, escuela o comunidad (en otras palabras, a través de *clusters*), pero el impacto se mide en los pacientes, alumnos o habitantes de la comunidad (véase el cuadro 15.1).[5] Cuando una evaluación de impacto genera *clusters*, es el número de estos últimos lo que determina en gran parte el tamaño de la muestra útil. En cambio, el número de individuos en los *clusters* importa menos. Se volverá sobre sobre esto más adelante.

La naturaleza de cualquier dato de la muestra construido a partir de programas que están conglomerados es algo diferente de las muestras obtenidas a partir de programas que no lo están. Como consecuencia, los cálculos de potencia comprenderán pasos ligeramente diferentes, dependiendo de si un programa asigna aleatoriamente los beneficios entre los *clusters* o sencillamente asigna los beneficios aleatoriamente entre todas las unidades de una población. Se analizará cada situación en su momento. Se comienza ahora con los principios de los cálculos de potencia en ausencia de *clusters*, es decir, cuando el tratamiento se asigna al nivel en que se observan los resultados. Luego se seguirá adelante para tratar los cálculos de potencia cuando hay *clusters*.

Cuadro 15.1 Ejemplos de *clusters*

Beneficio	Nivel al que se asignan los beneficios (*cluster*)	Unidad en que se miden los resultados
Transferencias monetarias	Pueblo	Hogares
Tratamiento anti malaria	Escuela	Individuos
Programa de capacitación	Barrio	Individuos

Cálculos de potencia sin *clusters*

Supóngase que se ha resuelto la primera pregunta estableciéndose que los beneficios del programa no se asignen por *cluster*. En otras palabras, el programa que se evalúa asigna de forma aleatoria los beneficios entre todas las unidades en una población elegible.

En el segundo paso, se deben identificar los *indicadores de resultado* más importantes para los cuales ha sido diseñado el programa. Estos indicadores derivan del objetivo del programa, de una teoría del cambio y de la pregunta fundamental de la investigación de la evaluación, como se señaló en la primera parte. Los cálculos de potencia también ayudarán a entender el tipo de indicadores más adecuados para las evaluaciones de impacto. En realidad, como se verá más adelante, se pueden requerir muestras de diversos tamaños para medir impactos en diferentes indicadores.

Tercero, se debe determinar el impacto mínimo que justificaría la inversión realizada en la intervención. Se trata sobre todo de una pregunta de políticas públicas, más que de una pregunta técnica. ¿Un programa de transferencias monetarias es una inversión provechosa si reduce la pobreza en 5%, 10% o 15%? ¿La implementación de un programa de mercado laboral activo vale la pena si aumenta los ingresos en 5%, 10% o 15%? La respuesta es sumamente específica del contexto, pero en todos los casos es necesario determinar el cambio en los indicadores de resultados que justificaría la inversión hecha en el programa. Dicho de otra manera, ¿cuál es el nivel de impacto por debajo del cual una intervención debería considerarse no exitosa? La respuesta a esa pregunta le dará el *efecto mínimo detectable* que la evaluación de impacto tiene que ser capaz de identificar. Responder a esta pregunta dependerá no solo del costo del programa y del tipo de beneficios que proporciona, sino también del costo de oportunidad de no invertir fondos en una intervención alternativa.

Si bien los efectos mínimos detectables se pueden basar en objetivos de políticas públicas, es posible utilizar otros enfoques para establecerlos. Puede que sea útil tomar como referencia efectos mínimos detectables en relación con resultados de los estudios en programas similares para arrojar luz sobre la magnitud de los impactos que se pueden esperar. Por ejemplo, las intervenciones en educación suelen medir los beneficios en términos de puntuaciones de las pruebas estandarizadas. Los estudios existentes demuestran que un aumento de 0,1 desviaciones típicas es relativamente pequeño, mientras que un aumento de 0,5 es relativamente grande. Como alternativa, se pueden llevar a cabo simulaciones ex ante para evaluar la gama de impactos que son realistas bajo diversas hipótesis. En el capítulo 1 se presentaron ejemplos de simulaciones ex ante para programas de transferencias monetarias condicionadas. Por último, los análisis económicos ex

Concepto clave

El *efecto mínimo detectable* (EMD) es el tamaño de efecto que una evaluación de impacto está diseñada para estimar para un determinado nivel de significancia y potencia. *Ceteris paribus*, se necesitan muestras más grandes para que una evaluación de impacto detecte diferencias más pequeñas entre los grupos de tratamiento y comparación o para detectar diferencias en un resultado más variable.

ante pueden arrojar luz sobre el tamaño de los impactos que se necesitarían para que la tasa de retorno de una determinada inversión sea suficientemente alta. Por ejemplo, los aumentos de los ingresos anualizados generados por un programa de capacitación laboral tendrían que ser superiores a la tasa de interés prevalente en el mercado.

Como se comprenderá, es más fácil identificar una gran diferencia entre dos grupos que identificar una diferencia pequeña. Para que una evaluación de impacto identifique una pequeña diferencia entre los grupos de tratamiento y comparación, se necesitará una estimación muy precisa de la diferencia de los resultados medios entre los dos grupos. Esto requiere una muestra grande. Como alternativa, en las intervenciones que se consideran viables solo si generan grandes cambios en los indicadores de resultado, las muestras necesarias para llevar a cabo una evaluación de impacto serán más pequeñas. Sin embargo, el efecto mínimo detectable debería fijarse de manera conservadora, dado que es menos probable que se detecte cualquier impacto menor que el efecto mínimo deseado.

Cuarto, para llevar a cabo cálculos de potencia, se le debe pedir a un experto que estime algunos parámetros básicos, como el promedio de la línea de base y una varianza de los indicadores de resultado. Estos valores de referencia deberían preferiblemente obtenerse de los datos recopilados en un contexto similar a aquel en el cual se implementará el programa que se estudia, o de una encuesta piloto en la población de interés.[6] Es muy importante señalar que cuanto más variables sean los resultados de interés, mayor será la muestra que se necesitará para estimar un efecto de tratamiento preciso. En el ejemplo de la intervención hipotética en nutrición, el peso de los niños es el resultado de interés. Si todos los individuos pesan lo mismo en la línea de base, será factible estimar el impacto de una intervención en nutrición en una muestra pequeña. En cambio, si los pesos de línea de base de los niños son muy variables, se requerirá una muestra más grande para estimar el impacto del programa.

Quinto, el equipo de evaluación tiene que determinar un *nivel de potencia* razonable y un *nivel de significancia* para la evaluación de impacto planificada. Como ya se señaló, la potencia de una prueba es igual a 1 menos la probabilidad de cualquier error de tipo II. Por lo tanto, la potencia oscila entre 0 y 1, donde un valor alto indica menos riesgo de no identificar un impacto existente. Una potencia de 0,8 es una referencia generalmente utilizada para los cálculos de potencia. Significa que se encontrará un impacto en el 80% de los casos allí donde se haya producido. Un nivel más alto de potencia de 0,9 (o 90%) a menudo proporciona una referencia útil pero más conservadora, lo cual aumenta el tamaño requerido de la muestra.

El nivel de significancia es la probabilidad de cometer un error de tipo I. Normalmente se fija en 5%, de modo que se puede tener una confianza

del 95% de llegar a la conclusión de que el programa ha tenido impacto si se encuentra un impacto significativo. Otros niveles habituales de significancia son 1% y 10%. Cuanto menor sea el nivel de significancia, más confianza se puede tener en que el impacto estimado es real.

Una vez que se han abordado estas cinco preguntas, el experto en cálculos de potencia puede calcular el tamaño requerido de la muestra utilizando un *software* estadístico.[7] El cálculo de potencia indicará el tamaño requerido de la muestra, dependiendo de los parámetros establecidos en los pasos 1 a 5. Los propios cálculos son sencillos, una vez que se han determinado los parámetros relevantes para las políticas (sobre todo en los pasos 2 y 3).[8] (Si a usted le interesa la implementación de los cálculos de potencia, el manual técnico disponible en el sitio web del libro contiene ejemplos de cálculos de potencia utilizando Stata y Optimal Design.)

Al solicitar asesoría a los expertos estadísticos, el equipo de evaluación debería pedir un análisis de la sensibilidad del cálculo de potencia ante cambios en los supuestos. Es decir, es importante entender cuánto tendrá que aumentar el tamaño requerido de la muestra con supuestos más conservadores (como un impacto previsto menor, mayor varianza en el indicador de resultado o un mayor nivel de potencia). También es una buena práctica encargar cálculos de potencia para diversos indicadores de resultados, dado que los tamaños requeridos de la muestra pueden variar considerablemente si algunos indicadores de resultados son mucho más variables que otros. Por último, los cálculos de potencia también pueden indicar el tamaño de la muestra necesario para establecer una comparación de los impactos del programa en diferentes subgrupos específicos (por ejemplo, hombres o mujeres, u otros subgrupos de la población de interés). Cada subgrupo tendría que tener el tamaño requerido de la muestra.

Evaluación del impacto del HISP: la decisión del tamaño de la muestra necesario para evaluar el HISP ampliado

Para volver al ejemplo presentado en la segunda parte del libro, supóngase que el ministerio de Salud estaba satisfecho con la calidad y los resultados de la evaluación del Programa de Subsidios de Seguros de Salud (HISP, por sus siglas en inglés). Sin embargo, antes de ampliar el programa, el ministro decide realizar una prueba piloto de una versión ampliada del programa, que denominan HISP+. El HISP original paga una parte del costo del seguro de salud de los hogares rurales pobres, y cubre los costos de la atención primaria y los medicamentos, pero no cubre la hospitalización. El ministro de Salud se pregunta si un HISP+ que

también cubra la hospitalización disminuiría aún más los gastos directos en salud de los hogares pobres. El ministerio le pide diseñar una evaluación de impacto para evaluar si el HISP+ disminuiría los gastos en salud de los hogares rurales pobres.

En este caso, elegir un diseño de evaluación de impacto no es difícil: el HISP+ tiene recursos limitados y no puede ser implementado universalmente de manera inmediata. Como consecuencia, se llega a la conclusión de que la asignación aleatoria sería el método de evaluación de impacto más viable y robusto. El ministro de Salud entiende que el método de asignación aleatoria puede funcionar bien y se muestra de acuerdo.

Para finalizar el diseño de la evaluación de impacto, usted contrata a un técnico estadístico que le ayudará a definir el tamaño de la muestra necesaria. Antes de comenzar a trabajar, el técnico estadístico le pide información clave. Utiliza una lista de verificación de cinco preguntas.

1. ¿El programa HISP+ generará *clusters*? A estas alturas, usted no está totalmente seguro. Cree que es posible asignar de forma aleatoria el paquete de beneficios ampliado a nivel de los hogares entre todos los hogares rurales pobres que ya se benefician del HISP. Sin embargo, usted sabe que el ministro de Salud puede preferir asignar el programa ampliado a nivel de la comunidad y que eso generaría *clusters*. El técnico estadístico sugiere llevar a cabo cálculos de potencia en un caso de referencia sin *clusters*, y luego analizar cómo cambiarían los resultados con los *clusters*.

2. ¿Cuál es el indicador de resultado? Usted explica que al gobierno le interesa un indicador bien definido, a saber, los gastos directos en salud de los hogares pobres. El técnico estadístico busca la fuente más actualizada para obtener valores de referencia de este indicador y sugiere utilizar la encuesta de seguimiento de la evaluación HISP. Señala que entre los hogares que reciben el HISP, los gastos directos anuales per cápita en salud llegan a un promedio de US$7,84.

3. ¿Cuál es el nivel mínimo de impacto que justificaría la inversión en la intervención? En otras palabras, ¿qué disminución de los gastos directos en salud por debajo del promedio de US$7,84 justificaría esta intervención? El profesional estadístico subraya que no se trata solo de una consideración técnica, sino de una decisión de políticas. Por eso, un responsable de las políticas como usted debe establecer el efecto mínimo que la evaluación debería ser capaz de detectar. Usted recuerda que, basándose en análisis económicos ex ante, el programa HISP+ se consideraría efectivo si redujera los gastos directos en salud de los

hogares en US$2. Aun así, usted sabe que para los fines de la evaluación, puede que sea preferible ser conservador al determinar el impacto mínimo detectable, dado que es poco probable que se detecte cualquier otro impacto menor. Para entender cómo el tamaño necesario de la muestra varía según el efecto mínimo detectable, usted sugiere que el técnico estadístico lleve a cabo cálculos para una reducción mínima de los gastos directos en salud de US$1, US$2 y US$3.

4. ¿Cuál es la varianza del indicador de resultado en la población de interés? El técnico estadístico vuelve a la base de datos de los hogares HISP tratados, y señala que la desviación típica de los gastos directos en salud es de US$8.

5. ¿Cuál sería un nivel razonable de potencia para la evaluación que se lleva a cabo? El profesional estadístico añade que los cálculos de potencia suelen efectuarse para una potencia de entre 0,8 y 0,9. Recomienda 0,9, pero propone realizar verificaciones de robustez más tarde, con un nivel menos conservador de 0,8.

Con toda esta información, el técnico estadístico emprende los cálculos de potencia. Como se había acordado, comienza con el caso más conservador de una potencia de 0,9. En el cuadro 15.2 se recogen los resultados que genera.

El estadístico llega a la conclusión de que para detectar una disminución de US$2 en los gastos directos en salud con una potencia de 0,9, la muestra tiene que contener al menos 672 unidades (336 unidades tratadas y 336 unidades de comparación, sin *clusters*). Señala que si usted se sintiera satisfecho detectando una disminución de US$3 en los gastos directos en salud, una muestra más pequeña de al menos 300 unidades (150 unidades en cada grupo) sería suficiente. En cambio, se necesitaría una

Cuadro 15.2 Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables, potencia = 0,9

Efecto mínimo detectable	Grupo de tratamiento	Grupo de comparación	Total muestra
US$1	1.344	1.344	2.688
US$2	336	336	672
US$3	150	150	300

Nota: El efecto mínimo detectable describe la reducción mínima de los gastos directos en salud de los hogares que puede detectar la evaluación de impacto. Potencia = 0,9; sin *clusters*.

muestra mucho más grande de al menos 2.688 unidades (1.344 cada grupo) para detectar una disminución de US$1 en los gastos directos en salud.

El técnico estadístico luego produce otro cuadro para un nivel de potencia de 0,8. El cuadro 15.3 muestra que los tamaños de la muestra requeridos son más pequeños con una potencia de 0,8 que con una potencia de 0,9. Para detectar una reducción de US$2 en los gastos directos en salud de los hogares, sería suficiente una muestra total de al menos 502 unidades. Para detectar una reducción de US$3, se precisan al menos 224 unidades. Sin embargo, para detectar una reducción de US$1 se necesitarían al menos 2.008 unidades en la muestra. El técnico estadístico subraya que los siguientes resultados son típicos de los cálculos de potencia:

- Cuanto mayor (más conservador) el nivel de potencia, mayor será el tamaño requerido de la muestra.
- Cuanto más pequeño el impacto detectado, mayor será el tamaño requerido de la muestra.

El técnico estadístico le pregunta si quiere llevar a cabo cálculos de potencia para otros resultados de interés. Usted sugiere considerar también el tamaño requerido de la muestra para detectar si el HISP+ influye en la tasa de hospitalización. En el ejemplo de las comunidades HISP tratadas, en el 5% de los hogares hay un miembro del hogar que acude al hospital en un año cualquiera; esto proporciona una tasa de referencia. El estadístico produce un nuevo cuadro, que demuestra que se necesitarían muestras relativamente grandes para detectar cambios en la tasa de hospitalización (cuadro 15.4) de 1, 2 o 3 puntos porcentuales con respecto a la tasa de línea de base del 5%.

Cuadro 15.3 Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables, potencia = 0,8

Efecto mínimo detectable	Grupo de tratamiento	Grupo de comparación	Total muestra
US$1	1.004	1.004	2.008
US$2	251	251	502
US$3	112	112	224

Nota: El efecto mínimo detectable describe la reducción mínima de los gastos directos en salud de los hogares que puede detectar la evaluación de impacto. Potencia = 0,8; sin *clusters*.

El cuadro 15.4 muestra que los requisitos del tamaño de la muestra son mayores para este resultado (la tasa de hospitalización) que para los gastos directos en salud. El técnico estadístico llega a la conclusión de que si usted está interesado en detectar impactos en ambos resultados, debería utilizar los tamaños de muestra más grandes que surgen de los cálculos de potencia efectuados en las tasas de hospitalización. Si se usan los tamaños de muestra de los cálculos de potencia realizados para los gastos directos, el técnico estadístico sugiere informar al ministro de Salud que la evaluación no tendrá suficiente poder para detectar efectos pertinentes para las políticas en las tasas de hospitalización.

Cuadro 15.4 Evaluación del HISP+: tamaño requerido de la muestra para detectar diversos efectos mínimos deseados (aumento de la tasa de hospitalización)
Potencia = 0,8; sin clusters

Efecto mínimo detectable (porcentaje)	Grupo de tratamiento	Grupo de comparación	Total muestra
1	7.257	7.257	14.514
2	1.815	1.815	3.630
3	807	807	1.614

Nota: El efecto mínimo deseado describe el cambio mínimo en la tasa de utilización de servicios hospitalarios (expresado en puntos porcentuales) que puede detectar la evaluación de impacto.

Pregunta HISP 8

A. ¿Qué tamaño de la muestra recomendaría usted para estimar el impacto del HISP+ en los gastos directos en salud?

B. ¿Ese tamaño de la muestra sería suficiente para detectar cambios en la tasa de hospitalización?

Cálculos de potencia con *clusters*

En el apartado anterior se introdujeron los principios de llevar a cabo cálculos de potencia para programas que no generan *clusters*. Sin embargo, como se señaló en la segunda parte, algunos programas asignan beneficios a nivel de *clusters*. A continuación, se describe brevemente cómo hay que adaptar los principios básicos de los cálculos de potencia para las muestras con *clusters*.

Ante la presencia de *clusters*, un principio rector clave es que el número de *clusters* suele importar mucho más que el número de individuos en los *clusters*.

Se requiere un número suficiente de *clusters* para probar de forma convincente si un programa ha tenido impacto al contraponer resultados en muestras de las unidades de tratamiento y comparación. Es el número de *clusters* el que determina en gran parte el tamaño de la muestra útil o efectivo. Si se asigna de manera aleatoria el tratamiento entre un pequeño número de *clusters*, es poco probable que los *clusters* de tratamiento y comparación sean idénticos. La asignación aleatoria entre dos distritos, dos escuelas o dos hospitales no garantizará que los dos *clusters* sean similares. En cambio, la asignación aleatoria de una intervención entre 100 distritos, 100 escuelas o 100 hospitales tiene más probabilidades de asegurar que los grupos de tratamiento y comparación sean similares. En resumen, se requiere un número suficiente de *clusters* para alcanzar un equilibrio. Además, el número de *clusters* también importa para la precisión de los efectos estimados del tratamiento. Se requiere un número suficiente de *clusters* para probar la hipótesis de que un programa tiene un impacto con suficiente potencia. Cuando se implementa una evaluación de impacto basada en la asignación aleatoria, es muy importante asegurar que el número de *clusters* sea suficientemente grande.

Se puede establecer el número de *clusters* requeridos para pruebas de hipótesis precisas efectuando cálculos de potencia. Esto exige formular las mismas cinco preguntas expuestas anteriormente, más una pregunta extra: ¿Cuán variable es el indicador de resultado en los *clusters*?

En el extremo, todos los resultados en un *cluster* están perfectamente correlacionados. Por ejemplo, puede ocurrir que el ingreso del hogar no varíe especialmente en las comunidades, pero que entre comunidades se observe una desigualdad importante en los ingresos. En este caso, si usted considera añadir una persona a su muestra de evaluación, agregar a un individuo de una comunidad nueva aumentará mucho más la potencia que introducir un individuo de una comunidad que ya está representada. Dado que los resultados están plenamente correlacionados en un *cluster*, añadir un nuevo individuo de ese *cluster* existente no aportará nueva información. En realidad, en este caso, es probable que el individuo de la segunda comunidad tenga un aspecto muy similar al individuo original ya incluido. En general, una mayor *correlación intra-cluster* en los resultados (es decir, una mayor correlación en los resultados o características entre las unidades que pertenecen al mismo *cluster*) aumenta el número de *clusters* requeridos para alcanzar un determinado nivel de potencia.

En las muestras con *clusters*, los cálculos de potencia subrayan los beneficios relativos entre añadir *clusters* y añadir observaciones dentro de los *clusters*. El aumento relativo de la potencia al agregar una unidad de un nuevo *cluster* es casi siempre mayor que el de sumar una unidad a un *cluster* ya existente. Aunque el incremento de la potencia al añadir un nuevo *cluster* puede ser drástico, agregar *clusters* también puede tener implicaciones

Concepto clave

El número de *clusters* importa mucho más en los cálculos de potencia que el número de individuos en los *clusters*. Se requieren a menudo al menos 40 a 50 *clusters* en cada uno de los grupos de tratamiento y comparación, aunque los requisitos del tamaño de la muestra variarán según los casos, y los cálculos de potencia son necesarios para asegurar un tamaño adecuado de la muestra.

operativas y elevar el costo de la implementación del programa o de la recopilación de datos. Más adelante en este capítulo, se explica cómo efectuar cálculos de potencia con *clusters* en el caso del HISP+ y se analizan algunas consideraciones.

En numerosos casos, se requieren al menos entre 40 y 50 *clusters* en cada grupo de tratamiento y comparación para obtener potencia suficiente y garantizar la similitud de las características de línea de base al usar métodos de asignación aleatoria. Sin embargo, puede que el número varíe de acuerdo con los diversos parámetros ya analizados, así como la correlación *intra-cluster*. Además, como se verá más adelante, es probable que el número probablemente aumente al utilizar métodos distintos de la asignación aleatoria (suponiendo que todos los demás factores permanezcan constantes).

Evaluación del impacto del HISP: tamaño requerido de la muestra para evaluar un HISP expandido con *clusters*

Después de su primera conversación con el técnico estadístico acerca de los cálculos de potencia para el HISP+, usted decide hablar brevemente con el ministro de Salud acerca de las implicaciones de asignar aleatoriamente los beneficios del HISP+ entre todos los individuos de la población que reciben el plan básico del HISP. La consulta revela que ese procedimiento no sería políticamente viable: en ese contexto, resultaría difícil explicar por qué una persona recibiría los beneficios ampliados mientras que su vecino no los recibiría.

Por lo tanto, en lugar de la asignación aleatoria a nivel individual, usted sugiere seleccionar aleatoriamente un cierto número de comunidades HISP para realizar una prueba piloto del HISP+. Todos los miembros de la comunidad del pueblo seleccionado serían elegibles. Este procedimiento generará *clusters* y, por lo tanto, requerirá nuevos cálculos de potencia. Ahora se trata de determinar el tamaño requerido de la muestra para evaluar el impacto del HISP+ cuando se asigne aleatoriamente por *cluster*.

Usted vuelve a consultar con su técnico estadístico. Él vuelve a asegurarle que solo se requiere un poco más de trabajo. En su lista de verificación solo queda una pregunta por responder, a saber: cuánto varía el indicador de resultado en los *clusters*. Por fortuna, también es una pregunta que se puede responder utilizando los datos del HISP. El técnico descubre que la correlación de los gastos directos en salud en la comunidad es igual a 0,04.

También pregunta si se ha fijado un límite para el número de comunidades en las que sería viable implementar el nuevo plan piloto. Dado que el programa ahora tiene 100 comunidades en el HISP, usted explica que podría tener, como máximo, 50 comunidades de tratamiento y 50 comunidades de comparación para el HISP+. Con esa información, el técnico estadístico produce los cálculos de potencia que aparecen en el cuadro 15.5 con una potencia de 0,8.

El estadístico llega a la conclusión de que para detectar una disminución de US$2 en los gastos directos en salud, la muestra debe incluir al menos 630 unidades, es decir, 7 unidades por *cluster* en 90 *clusters* (45 en el grupo de tratamiento y 45 en el grupo de comparación). Señala que este número es mayor que en la muestra con asignación aleatoria a nivel de los hogares, que requirió solo un total de 502 unidades (251 en el grupo de tratamiento y 251 en el grupo de comparación; véase el cuadro 15.3). Para detectar una disminución de US$3 en los gastos directos en salud, la muestra tendría que incluir al menos 246 unidades, o 3 unidades en cada uno de los 82 *clusters* (41 en el grupo de tratamiento y 41 en el grupo de comparación).

Posteriormente, el técnico estadístico le indica cómo el número total de observaciones requeridas en la muestra varía con el número total de *clusters*. Luego decide repetir los cálculos para un efecto mínimo detectable de US$2 y una potencia de 0,8. El tamaño de la muestra total requerida para estimar dicho efecto aumenta visiblemente cuando el número de *clusters* disminuye (cuadro 15.6). Con 120 *clusters*, se necesitaría una muestra de 600 observaciones. Si solo hubiera 30 *clusters* disponibles, la muestra total debería contener 1.500 observaciones. En cambio, si hubiera 90 *clusters* disponibles, solo se necesitarían 630 observaciones.

Cuadro 15.5 Evaluación del HISP+: tamaño requerido de la muestra para identificar diversos efectos mínimos detectables (disminución de los gastos del hogar en salud)
Potencia = 0,8; máximo de 100 clusters

Efecto mínimo detectable	Número de *clusters*	Unidades por *cluster*	Total muestra con *clusters*	Total muestra sin *clusters*
US$1	100	102	10.200	2.008
US$$2	90	7	630	502
US$3	82	3	246	224

Nota: El efecto mínimo detectable describe la reducción mínima de los gastos directos en salud de los hogares que puede detectar la evaluación de impacto. El número de *clusters* es el número total de *clusters*, la mitad de los cuales será el número de *clusters* del grupo de comparación, y la otra mitad será el número de *clusters* del grupo de tratamiento.

Cuadro 15.6 Evaluación del HISP+: tamaño requerido de la muestra para detectar un impacto mínimo de US$2 en diversas cantidades de *clusters*
Potencia = 0,8

Efecto mínimo detectable	Número de *clusters*	Unidades por *cluster*	Total muestra con *clusters*
US$2	30	50	1.500
US$2	58	13	754
US$2	81	8	648
US$2	90	7	630
US$2	120	5	600

Nota: El número de *clusters* es el número total de *clusters*, la mitad de los cuales será el número de *clusters* del grupo de comparación, y la otra mitad será el número de *clusters* del grupo de tratamiento. Si el diseño no tuviera *clusters*, se necesitarían 251 unidades en cada grupo para identificar un efecto mínimo detectable de US$2 (véase el cuadro 15.3).

Pregunta HISP 9

A. ¿Qué tamaño total de la muestra recomendaría para estimar el impacto del HISP+ en los gastos directos en salud?

B. ¿En cuántas comunidades le aconsejaría al ministro de Salud implementar el HISP+?

Más allá del caso de referencia

Este capítulo se ha centrado en el caso de referencia de una evaluación de impacto implementada utilizando el método de asignación aleatoria, con pleno cumplimiento. Este es el escenario más sencillo y, por lo tanto, el más adecuado para transmitir la intuición en que se basan los cálculos de potencia. Aun así, numerosos aspectos prácticos de nuestros cálculos de potencia aún no han sido analizados, y es necesario considerar detenidamente las desviaciones de los casos básicos que se abordan aquí. Más abajo, se tratan algunas de estas desviaciones.

Utilización de métodos cuasi experimentales. Ceteris paribus, los métodos de evaluación de impacto cuasi experimentales, como la regresión discontinua, el pareamiento o las diferencias en diferencias, tienden a requerir muestras más grandes que el método de referencia de asignación aleatoria. Por ejemplo, al utilizar el diseño de regresión discontinua, en el capítulo 6 se subrayaba que solo se pueden considerar las observaciones en torno al umbral de elegibilidad. Se requiere una muestra suficientemente

grande en torno a ese umbral. Los cálculos de potencia son necesarios para estimar la muestra requerida de modo de establecer comparaciones significativas en torno al umbral.

Por otro lado, la disponibilidad de diversas rondas de datos puede contribuir a aumentar la potencia de una evaluación de impacto con un determinado tamaño de la muestra. Por ejemplo, los datos de línea de base sobre resultados y otras características pueden añadir precisión a la estimación de los efectos de tratamiento. La disponibilidad de medidas repetidas de resultados después del comienzo del tratamiento también puede ser útil.

Análisis de diferentes modalidades de programa o innovaciones de diseño. En los ejemplos presentados en este capítulo, el tamaño total de la muestra se dividía por igual entre los grupos de tratamiento y comparación. En algunos casos, la principal pregunta de políticas con respecto a la evaluación puede generar la comparación de impactos del programa entre las modalidades del programa o las innovaciones de diseño. Si esto es así, el impacto previsto puede ser relativamente menor que si un grupo de tratamiento objeto de un programa fuera comparado con un grupo de comparación que no recibía ningún tipo de beneficios. Como tal, el efecto mínimo deseado entre los dos grupos de tratamiento puede ser más pequeño que el efecto mínimo deseado entre el grupo de tratamiento y el grupo de comparación. Esto implicaría que la distribución óptima de la muestra generaría grupos de tratamiento que son relativamente más grandes que el grupo de comparación.[9] En las evaluaciones de impacto con múltiples ramas de tratamiento, puede que sea necesario implementar cálculos de potencia para estimar por separado el tamaño de cada grupo de tratamiento y comparación, en función de la principal pregunta de interés de las políticas.

Comparación de subgrupos. En otros casos, algunas de las preguntas de la evaluación de impacto pueden centrarse en estimar si los impactos de un programa varían entre diferentes subgrupos, como el sexo, la edad o las categorías de ingreso. Si esto es lo que ocurre, los requisitos del tamaño de la muestra serán mayores y los cálculos de potencia tendrán que ajustarse de forma correspondiente. Por ejemplo, una pregunta clave de políticas puede ser si un programa educativo tiene un impacto mayor en las alumnas que en los alumnos. Se necesitará un número suficiente de alumnos de cada sexo en el grupo de tratamiento y el grupo de comparación para detectar un impacto en cada subgrupo. Si se pretende comparar los impactos del programa entre dos subgrupos, puede que se duplique el tamaño requerido de la muestra. Si se considera la heterogeneidad entre más grupos (por ejemplo, por la edad) también puede aumentar considerablemente el tamaño requerido de la muestra. Si este tipo de comparaciones entre grupos ha de llevarse a cabo en el contexto de una evaluación de impacto que depende de la asignación

aleatoria, es preferible también tenerlas en cuenta cuando se implementa la aleatorización y, sobre todo, para aplicar una asignación aleatoria por bloques o estratos (es decir, en cada subgrupo que se compara). En la práctica, aunque no se realice ninguna comparación entre subgrupos, la aleatorización estratificada o por bloque puede contribuir a maximizar aún más la potencia de un determinado tamaño de la muestra.

Análisis de múltiples resultados. Es necesario proceder con singular cuidado cuando se emprenden cálculos de potencia en los casos en que una evaluación de impacto pretenda probar si un programa genera cambios en múltiples resultados. Si se tienen en cuenta numerosos resultados diferentes, habrá una probabilidad relativamente más alta de que la evaluación de impacto encuentre impactos en uno de los resultados solo por azar. Para abordar esto, el equipo de evaluación de impacto tendrá que pensar en probar la significancia estadística conjunta de los cambios en diversos resultados. Como alternativa, se pueden elaborar algunos índices o familias de resultados. Estos enfoques para lidiar con las pruebas de múltiples hipótesis tienen implicaciones para los cálculos de potencia y el tamaño de la muestra y, en ese sentido, hay que tenerlos en cuenta cuando se define la muestra necesaria para la evaluación de impacto.[10]

Para lidiar con el cumplimiento imperfecto o el desgaste de la muestra. Los cálculos de potencia suelen proporcionar el tamaño mínimo requerido de la muestra. En la práctica, los problemas de implementación a menudo implican que el tamaño de la muestra real es más pequeño que el tamaño planificado. Por ejemplo, el cumplimiento imperfecto puede significar que solo se inscribe una parte de los beneficiarios a los que se ofrece el programa. Los requisitos del tamaño de la muestra aumentan cuando surge el cumplimiento imperfecto. Además, aunque todos los individuos se inscribieran en el programa, se puede producir algún grado de desgaste en la encuesta de seguimiento si no se da con el paradero de todos los individuos. Aunque ese incumplimiento o desgaste es aleatorio y no afecta la consistencia de las estimaciones de impacto, estos aspectos influirían en la potencia de la evaluación de impacto. Para dar cuenta de dichos factores, generalmente se recomienda añadir un margen al tamaño de la muestra prevista por los cálculos de potencia. De la misma manera, los datos de menor calidad tendrán más error de medición y harán que los resultados de interés sean más variables, además de que requerirán tamaños de la muestra más grandes.

Las reflexiones más avanzadas mencionadas en esta sección exceden el alcance de este libro, pero los recursos recogidos al final de este capítulo pueden ser útiles. En la práctica, los equipos de evaluación tienen que incluir o contratar a un experto que pueda efectuar cálculos de potencia, y el experto debería ser capaz de asesorar en temas más complejos.

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para ejemplos de cómo efectuar cálculos de potencia con los programas Stata™ y Optimal Design para el caso específico del HISP que ilustra este capítulo, véase el manual técnico disponible en el sitio web del libro (http://www.iadb.org/portalevaluacion). Este manual incluye material técnico adicional para lectores con conocimientos de estadística y econometría.
- Para un debate detallado sobre el muestreo (incluyendo otros métodos como el muestreo sistemático o muestreo de múltiples etapas), más allá de los conceptos básicos tratados aquí, véase los siguientes recursos:
  - W. G. Cochran (1977), *Sampling Techniques,* tercera edición. Nueva York: John Wiley.
  - L. Kish (1995), *Survey Sampling.* Nueva York: John Wiley.
  - S. Lohr (1999), *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks Cole.
  - S. K. Thompson (2002), *Sampling,* segunda edición. Nueva York: John Wiley.
  - O, en un nivel más básico, G. Kalton (1983), *Introduction to Survey Sampling.* Beverly Hills, CA: Sage Publications.
- Se puede encontrar orientación práctica para el muestreo en:
  - M. Grosh y J. Muñoz (1996), "A Manual for Planning and Implementing the Living Standards Measurement Study Survey." Documento de trabajo LSMS 126. Washington, D.C.: Banco Mundial.
  - Naciones Unidas (2005), *Household Sample Surveys in Developing and Transition Countries.* Nueva York: Naciones Unidas.
  - G. Iarossi (2006), *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, D.C.: Banco Mundial.
  - A. G. Fink (2008), *How to Conduct Surveys: A Step by Step Guide*, cuarta edición. Beverly Hills, CA: Sage.
- Para una hoja de cálculo de potencia que calcule la potencia de un determinado tamaño de la muestra después de ingresar ciertas características, véase el portal de evaluación del Banco Interamericano de Desarrollo, sección de diseño, en herramientas (http://www.iadb.org/portalevaluacion).
- Para más información sobre cálculos de potencia y tamaño de la muestra, véase el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial, Módulo 3 sobre Diseño. Este módulo también incluye un guía para realizar cálculos de potencia ex ante, un documento sobre cálculos de potencia con variables binarias y una recopilación de referencias útiles para más información sobre los cálculos de potencia (http://www.worldbank.org/health/impactevaluationtoolkit).
- Para diversos blogs sobre cálculos de potencia, véase el blog de impacto del desarrollo del Banco Mundial (*World Bank Development Impact Blog*) (http://blogs.worldbank.org/impactevaluations/).

- Para un debate de algunas reflexiones sobre cálculos de potencia en diseños más complejos que el caso de referencia de la asignación aleatoria en presencia de cumplimiento perfecto, véase:
  - ➢ J. Spybrook, S. Raudenbush, X. Liu, R. Congdon y A. Martínez (2008), *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. Nueva York: William T. Grant Foundation.
  - ➢ P. Rosenbaum (2009), "The Power of Sensitivity Analysis and Its Limit." En: P. Rosenbaum, *Design of Observational Studies*, capítulo 14. Nueva York: Springer Series in Statistics.
- Sobre el tema de pruebas de múltiples hipótesis, véase:
  - ➢ E. Duflo, R. Glennerster, M. Kremer, T. P. Schultz y A. S. John (2007), "Using Randomization in Development Economics Research: A Toolkit." En: *Handbook of Development Economics,* Vol. 4, capítulo 61, pp. 3895–3962. Ámsterdam: Elsevier.
  - ➢ P. Z. Schochet (2008), *Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions*. Preparado por Mathematica Policy Research Inc., para el Institute of Education Sciences, U.S. Department of Education, Washington, D.C.
- Hay diversos instrumentos disponibles para quienes estén interesados en analizar el diseño de la muestra de manera más profunda. Por ejemplo, la W.T. Grant Foundation desarrolló el *software* de libre disponibilidad Optimal Design, un *software* para la Investigación de Múltiples Niveles y Longitudinal, útil para el análisis de potencia estadística con *clusters*. El *software* y el manual de Optimal Design se pueden descargar de http://hlmsoft.net/od.

Notas

1. Estrictamente hablando, las muestras se extraen de los marcos muestrales. En este análisis, se supone que el marco muestral coincide perfectamente con la población.
2. Como se señaló en la segunda parte, el cumplimiento supone que todas las unidades asignadas al grupo de tratamiento son tratadas y que todas las unidades asignadas al grupo de comparación no son tratadas.
3. En este contexto, el término *población* no se refiere a la población del país, sino al conjunto del grupo de niños que nos interesan: la población de interés.
4. Esta idea se concreta en el teorema denominado "teorema del límite central". En términos formales, en el caso de un resultado *y*, el teorema del límite central establece que la media de la muestra $\overline{y}$ constituye en promedio una estimación válida de la media de la población. Además, para un tamaño de muestra *n* y una varianza de la población σ^2, la varianza de la media de la muestra es inversamente proporcional al tamaño de la muestra:

$$var(\overline{y}) = \frac{\sigma^2}{n}$$

A medida que aumenta el tamaño de la muestra *n*, las estimaciones de la varianza de la muestra tienden hacia cero. En otras palabras, la media se estima con más precisión en muestras grandes que en pequeñas.

5. Las cuestiones de índole social y política, que hacen imposible la asignación aleatoria en los *clusters* suelen requerir la asignación de beneficios por *cluster*. En el contexto de una evaluación de impacto, la configuración de *clusters* suele ser necesaria debido a los probables efectos de derrame, o a la contaminación de los beneficios del programa entre los individuos en los *clusters*. Véase el tratamiento de este tema en el capítulo 11.
6. Cuando se calcula la potencia desde la línea de base, la correlación entre los resultados a lo largo del tiempo también se debe considerar en el cálculo de potencia.
7. Por ejemplo, Spybrook et al. (2008) introdujeron el Optimal Design, un programa informático fácil de usar para efectuar cálculos de potencia.
8. En general, es deseable contar con grupos de tratamiento y de comparación del mismo tamaño. De hecho, para cierto número de observaciones en una muestra, se maximiza la potencia asignando la mitad de las observaciones al grupo de tratamiento y la otra mitad al grupo de comparación. Sin embargo, los grupos de tratamiento y de comparación no siempre tienen que tener el mismo tamaño.
9. Los costos del tratamiento también se pueden tener en cuenta y generar grupos de tratamiento y comparación que no tienen el mismo tamaño. Véase, por ejemplo Duflo et al. (2007).
10. Véase, por ejemplo Duflo et al. (2007) o Schochet (2008).

Referencias bibliográficas

Cochran, W. G. 1977. *Sampling Techniques*, tercera edición. Nueva York: John Wiley & Sons.

Duflo, E., R. Glennerster, M. Kremer et al. 2007. "Using Randomization in Development Economics Research: A Toolkit." En: T. P. Schultz y J. Strauss (eds.), *Handbook of Development Economics,* Vol. 4, pp. 3895–962. Ámsterdam: Elsevier.

Fink, A. G. 2008. *How to Conduct Surveys: A Step by Step Guide*, cuarta edición. Beverly Hills, CA: Sage.

Grosh, M. y P. Glewwe (eds.). 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study.* Washington, D.C.: Banco Mundial.

Grosh, M. y J. Muñoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Survey." Documento de trabajo LSMS 126. Washington, D.C.: Banco Mundial.

Iarossi, G. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, D.C.: Banco Mundial.

Kalton, G. 1983. *Introduction to Survey Sampling.* Beverly Hills, CA: Sage.

Kish, L. 1995. *Survey Sampling*. Nueva York: John Wiley.

Lohr, S. 1999. *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks Cole.

Rosenbaum, P. 2009. *Design of Observational Studies*. Nueva York: Springer Series in Statistics.

Schochet, P. Z. 2008. *Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions*. NCEE 2008-4018. National Center for Eduational Evaluation and Regional Assistance, Institute of Education Sciences. Washington, D.C.: U.S. Department of Education.

Spybrook, J., S. Raudenbush, X. Liu, R. Congdon y A. Martínez. 2008. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. Nueva York: William T. Grant Foundation.

Thompson, S. K. 2002. *Sampling*, segunda edición. Nueva York: John Wiley.

Vermeersch, C., E. Rothenbühler y J. Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. Washington, D.C.: Banco Mundial. Disponible en http://www.worldbank.org/health/impactevaluationtoolkit.

Capítulo 16

Encontrando fuentes adecuadas de datos

Tipos de datos necesarios

En este capítulo se analizan las diversas fuentes de datos que pueden utilizar las evaluaciones de impacto. En primer lugar, se estudian las fuentes de datos existentes, sobre todo los datos administrativos, y se proporcionan algunos ejemplos de evaluaciones de impacto que han aprovechado datos existentes. Dado que muchas evaluaciones requieren la recopilación de datos nuevos, también se examinan los pasos en la recopilación de nuevos datos de las encuestas. Una comprensión clara de estos pasos contribuirá a asegurar que la evaluación de impacto se base en datos de calidad que no comprometan el diseño de evaluación. Como primer paso, habrá que contratar la elaboración de un cuestionario adecuado. Paralelamente, se necesitará ayuda de una empresa o un organismo del gobierno especializado en recopilación de datos. La entidad de recopilación de datos reclutará y capacitará al personal de campo y realizará una prueba piloto del cuestionario. Después de introducir los ajustes necesarios, la empresa o el organismo podrán proceder con el trabajo de campo, recopilar los datos, digitalizarlos y procesarlos antes de que puedan ser entregados, almacenados y analizados por el equipo de evaluación.

Para evaluar el impacto de la intervención en los resultados de interés, se requieren datos de buena calidad. La cadena de resultados que se expone en

el capítulo 2 proporciona una base para definir qué indicadores deberían medirse y cuándo. Los indicadores son necesarios en toda la cadena de resultados.

Datos sobre los resultados. La primera y principal necesidad son los datos sobre los indicadores de resultado directamente afectados por el programa. Los indicadores de resultado están vinculados con los objetivos que el programa pretende alcanzar. Como se señaló en el capítulo 2, los indicadores de resultado deben seleccionarse preferiblemente para que sean específicos, medibles, atribuibles, realistas y focalizados (EMARF). No obstante, la evaluación de impacto no debe medir solo aquellos resultados de los que el programa rinde directamente cuentas. Los datos sobre los indicadores de resultados que el programa afecta indirectamente, o los indicadores que capturan los efectos no intencionados del programa, maximizarán el valor de la información que genera la evaluación de impacto, así como la comprensión de la efectividad general del programa.

Datos sobre los resultados intermedios. Por otro lado, los datos sobre los resultados intermedios son útiles para ayudar a entender los canales a través de los cuales el programa evaluado ha tenido impacto –o no lo ha tenido– en los resultados finales de interés. Normalmente, las evaluaciones de impacto se llevan a cabo a lo largo de diversos períodos y se debe definir cuándo medir los indicadores de resultados. Siguiendo esta cadena de resultados, se puede establecer una jerarquía de indicadores de resultado, que abarca desde los indicadores de corto plazo, que se pueden medir mientras los participantes todavía están en el programa, como la asistencia escolar registrada en una encuesta de seguimiento de corto plazo en el contexto de un programa educativo, hasta las encuestas de seguimiento de más largo plazo, como el aprendizaje escolar o la inserción en el mercado laboral, que se pueden medir en una encuesta de seguimiento a más largo plazo después de que los participantes han dejado el programa. Para medir el impacto a lo largo del tiempo de manera convincente es necesario contar con datos de la línea de base antes de implementar el programa o la innovación que se evalúa. La sección del capítulo 12 que versa sobre la programación en el tiempo de la evaluación arroja luz sobre cómo definir el momento en que se recopilan los datos.

Concepto clave

Los indicadores son necesarios en toda la cadena de resultados. Constituyen la verificación para medir los resultados finales y los resultados intermedios, así como los beneficios y la calidad de implementación del programa.

Como se señaló en el capítulo 15, en el contexto de los cálculos de potencia, cuando las muestras son relativamente pequeñas, algunos indicadores pueden no ser adecuados para la evaluación de impacto. Detectar el impacto de una intervención cuyos indicadores son extremadamente variables, se refieren a sucesos poco frecuentes o solo se ven afectados de forma marginal por la intervención, puede requerir muestras demasiado grandes. Por ejemplo, solo será posible determinar el impacto de una intervención sobre las tasas de mortalidad materna si se dispone de una muestra de decenas de

miles de mujeres embarazadas, dado que la mortalidad es (afortunadamente) un hecho excepcional. En ese caso, puede que sea necesario replantear la evaluación de impacto y focalizarla en indicadores más intermedios, relacionados con los resultados finales, pero para los cuales hay suficiente potencia como para detectar efectos. En el caso de una intervención cuyo fin es reducir la mortalidad materna, un indicador intermedio podría estar vinculado con la utilización de los servicios de salud durante el embarazo, y con los partos en los centros de salud, que están asociados con la mortalidad. Los cálculos de potencia analizados en el capítulo 15 pueden contribuir a arrojar luz sobre los indicadores en los que se detectan impactos y aquellos en los que puede ser más difícil detectar impactos sin muestras muy grandes.

Datos sobre las actividades y productos del programa. También se requieren indicadores para la parte de la cadena de resultados que describe las actividades y productos del programa. Concretamente, los *datos de monitoreo del programa* pueden proporcionar información esencial sobre las prestaciones de la intervención. En particular, los datos de monitoreo incluyen definiciones sobre quiénes son los beneficiarios y qué beneficios o productos del programa pueden haber recibido. Como mínimo, se necesitan datos de monitoreo para saber cuándo comienza un programa y quién recibe beneficios, así como para proporcionar una medida de la intensidad o calidad de la intervención. Esto es particularmente importante en los casos en que un programa puede no llegar a todos los beneficiarios con el mismo contenido, calidad o duración. Es esencial tener una comprensión adecuada de la medida en que la intervención se ha implementado siguiendo el diseño, para interpretar los resultados de la evaluación de impacto, lo que incluye saber si destacan la efectividad del programa puesto en marcha según el diseño o si hay deficiencias en su implementación.

Datos adicionales. Puede que se precisen otros datos para la evaluación de impacto, lo cual depende de la metodología usada. Los datos sobre otros factores que pueden influir en el resultado de interés pueden ser necesarios para controlar por influencias externas. Este aspecto es particularmente importante cuando se utilizan métodos de evaluación que dependen de más supuestos que los métodos aleatorios. A veces también es necesario tener datos sobre los resultados y otros factores a lo largo del tiempo para calcular tendencias, como sucede con el método de diferencias en diferencias. Dar cuenta de otros factores y tendencias anteriores también contribuye a aumentar la potencia estadística. Incluso con la asignación aleatoria, los datos sobre otras características pueden ayudar a estimar los efectos del tratamiento con más precisión. También pueden ser utilizados para incluir controles adicionales o analizar la heterogeneidad de los efectos del programa en características relevantes.

El diseño seleccionado para la evaluación de impacto también afectará a los requisitos de datos. Por ejemplo, si se elige el método de pareamiento o de diferencias en diferencias, habrá que recolectar datos sobre una gama muy amplia de características para los grupos tanto de tratamiento como de comparación, lo que hace posible ejecutar una serie de pruebas de robustez, como se explicó en la segunda parte o en el capítulo 11 (véase el cuadro 11.2).

Para cada evaluación, resulta útil desarrollar una matriz que enumere las preguntas de interés, los indicadores de resultado para cada pregunta y la fuente de los datos, como se describe en el gráfico 2.1 del capítulo 2 sobre la cadena de resultados. La elaboración de un plan de evaluación de impacto y el contar con un plan de preanálisis constituyen otras oportunidades esenciales para definir una lista precisa de indicadores clave requeridos en las evaluaciones de impacto.

La utilización de datos cuantitativos existentes

Una de las primeras cuestiones que se debe considerar cuando se diseña una evaluación de impacto es qué fuente de datos se utilizará. Una consideración fundamental es si la evaluación de impacto dependerá de datos existentes o si requerirá la recopilación de datos nuevos.

Casi siempre se necesitan datos existentes al comienzo de una evaluación de impacto para estimar los valores de referencia de los indicadores o para efectuar cálculos de potencia, como se analizó en el capítulo 15. Después de la fase de planificación, la disponibilidad de datos existentes puede disminuir de forma considerable el costo de una evaluación de impacto. Si bien es probable que los datos existentes y, en particular los datos administrativos, sean subutilizados en la evaluación de impacto en general, la viabilidad de usar datos existentes en la evaluación de impacto tiene que ser valorada con detenimiento.

De hecho, como se señaló en el capítulo 12, la recopilación de datos suele representar el mayor costo de una evaluación de impacto. Sin embargo, para determinar si los datos existentes se pueden utilizar en una determinada evaluación de impacto, debe tenerse en cuenta una serie de preguntas:

- El *muestreo*. ¿Se dispone de datos existentes tanto para el grupo de tratamiento como para el grupo de comparación? ¿Las muestras existentes se han extraído de un marco muestral que coincide con la población de interés? Las unidades del marco muestral, ¿se han obtenido mediante un procedimiento de muestreo probabilístico?

- *Tamaño de la muestra*. ¿Las series de datos son suficientemente grandes para detectar cambios en los indicadores de resultado con suficiente

potencia? La respuesta a esta pregunta depende de la elección de los indicadores de resultado, así como de los resultados de los cálculos de potencia tratados en el capítulo 15.

- *Disponibilidad de datos de línea de base*. ¿Los datos existentes disponibles tanto para los grupos de tratamiento como de comparación son anteriores a la implementación del programa o de la innovación que se evalúa? La disponibilidad de datos en línea de base es esencial para documentar el equilibrio en las características previas del programa entre los grupos de tratamiento y de comparación cuando se utilizan métodos aleatorios, y son esenciales para la implementación de diseños cuasi experimentales.
- *Frecuencia*. ¿Los datos existentes son recopilados con suficiente frecuencia? ¿Se dispone de ellos para todas las unidades de la muestra a lo largo del tiempo, incluyendo los momentos en que hay que medir los indicadores de resultado según la cadena de resultados y la lógica de la intervención?
- *Alcance*. ¿Los datos existentes contienen todos los indicadores necesarios para responder a las preguntas de interés de las políticas, incluyendo los principales indicadores de resultado y los resultados intermedios de interés?
- *Vínculos con la información de monitoreo del programa*. ¿Los datos existentes se pueden vincular a datos del monitoreo de la implementación del programa, lo que implica observar qué unidades pertenecen a los grupos de tratamiento y de comparación, y si todas las unidades asignadas al grupo de tratamiento reciben los mismos beneficios?
- *Identificadores únicos*. ¿Existen identificadores únicos que vinculen diferentes fuentes de datos?

Como lo subrayan las preguntas anteriores, los requisitos para los datos existentes son bastante importantes, y no es habitual que los datos existentes resulten suficientes para las evaluaciones de impacto. Aun así, con el rápido crecimiento en el alcance y la cobertura de los sistemas de información, y con la evolución general hacia un mundo en que los datos digitales de una amplia gama de fuentes se almacenan de manera periódica, cada vez más evaluaciones de impacto pueden contemplar el uso de datos existentes. Se puede utilizar una gama de fuentes potenciales de datos existentes en las evaluaciones de impacto, lo cual abarca datos censales, encuestas nacionales o datos administrativos.

Los *datos del censo de población* pueden proporcionar información exhaustiva sobre toda la población. Se pueden utilizar para las evaluaciones de impacto cuando están disponibles en un nivel suficientemente

desagregado e incluyen detalles para saber qué unidades pertenecen al grupo de tratamiento o de comparación, como los identificadores geográficos o personales. Los datos censales no se recopilan a menudo, y normalmente incluyen solo un pequeño conjunto de indicadores clave. Sin embargo, en ocasiones dichos datos se recopilan para incluirse en sistemas de información o registros que proporcionan la base para definir los objetivos de los programas públicos, lo que incluye identificadores únicos que pueden servir de soporte a vínculos con otras bases de datos existentes.

Las *encuestas representativas a nivel nacional*, como las encuestas de hogares, las encuestas de medición de los niveles de vida, las encuestas de la fuerza laboral, las encuestas demográficas y de salud, las encuestas de empresas o las encuestas de instalaciones también se pueden contemplar. Estas pueden contener un conjunto exhaustivo de variables de resultado, pero rara vez cuentan con suficientes observaciones, tanto del grupo de tratamiento como de comparación, para llevar a cabo una evaluación de impacto. Supóngase, por ejemplo, que se desea evaluar un programa nacional de gran alcance que llega al 10% de los hogares en un determinado país. Si una encuesta representativa a nivel nacional se lleva a cabo en 5.000 hogares cada año, esta puede contener alrededor de 500 hogares que reciben el programa en cuestión. ¿Es la muestra lo bastante grande para llevar a cabo una evaluación de impacto? Los cálculos de potencia pueden responder a esta pregunta, pero en muchos casos la respuesta es negativa.

Además de determinar si se pueden utilizar las encuestas existentes, también se debe averiguar si se están planificando nuevas iniciativas de recopilación de datos nacionales. Si se planifica una encuesta que cubrirá la población de interés, quizá también se pueda introducir una pregunta o una serie de preguntas como parte de esa encuesta. Si ya se ha planeado una encuesta que mide los indicadores requeridos, existe la posibilidad de sobre muestrear una determinada población para asegurar una cobertura adecuada en los grupos de tratamiento y comparación y acomodar la evaluación de impacto. Por ejemplo, la evaluación del Fondo Social de Nicaragua complementó un estudio nacional de medición de los niveles de vida con una muestra adicional de beneficiarios (Pradhan y Rawlings, 2002).

Los *datos administrativos* suelen ser recopilados por organismos públicos o agencias privadas como parte de sus operaciones regulares, normalmente con cierta frecuencia, y a menudo para monitorear los servicios prestados o registrar interacciones con los usuarios. En algunos casos, los datos administrativos contienen los indicadores de resultado necesarios para una evaluación de impacto. Por ejemplo, los sistemas educativos cuentan con registros de la matriculación y asistencia de los alumnos y de las calificaciones de las pruebas, y también pueden recopilar información sobre los

insumos escolares y los maestros. De la misma manera, los sistemas de salud pueden reunir datos sobre las características y localización de los centros de salud, la oferta de servicios de salud y la asignación de recursos. También pueden consolidar datos recopilados en centros de salud sobre los historiales médicos de los pacientes, datos antropométricos, historiales de vacunaciones y, de manera más amplia, datos sobre la incidencia de las enfermedades y estadísticas vitales. Las empresas de servicios públicos reúnen datos sobre el consumo de agua o electricidad. Las agencias tributarias pueden recoger datos sobre los ingresos y los impuestos. Los sistemas de transporte recopilan datos sobre los pasajeros y los tiempos de viaje. Las empresas del sistema financiero recopilan datos sobre las transacciones o el historial crediticio de los clientes. Todas estas fuentes de datos existentes pueden ser potencialmente utilizadas en las evaluaciones de impacto. A veces incluyen series temporales extensas, que pueden contribuir a seguir a las unidades a lo largo del tiempo.

Es crucial realizar un diagnóstico de la disponibilidad y calidad de los datos cuando se considera la posibilidad de utilizar datos administrativos. En algunos casos, los datos de las fuentes administrativas pueden ser más fiables que los datos de las encuestas. Por ejemplo, un estudio en Malawi reveló que los encuestados daban información falsa sobre la asistencia y matriculación escolar en una encuesta de hogares, en comparación con los registros administrativos obtenidos en las escuelas; por lo tanto, los resultados de la evaluación de impacto eran más fiables si se basaban en los datos administrativos (Baird y Özler, 2012). Al mismo tiempo, en numerosos contextos, los datos administrativos son recopilados por un gran número de proveedores y pueden ser de calidad desigual. Por lo tanto, su fiabilidad debe valorarse detenidamente antes de tomar la decisión de trabajar con datos administrativos en la evaluación de impacto. Un aspecto crucial consiste en asegurar que existan identificadores únicos para vincular los datos administrativos con otras fuentes de datos, incluyendo datos sobre el monitoreo del programa que documentan qué unidades han recibido los beneficios del programa. Cuando estos identificadores existen, como los números de identificación nacional usados de manera consistente, se puede evitar una gran cantidad de trabajo para preparar y limpiar los datos. En todos los casos, la protección de la confidencialidad es una parte importante de la preparación de los datos y del protocolo de gestión de datos. Los principios éticos que rigen la protección de sujetos humanos (véase el debate en el capítulo 13) también rigen el uso que se hace de datos existentes.

Algunas evaluaciones retrospectivas influyentes han trabajado con registros administrativos: Galiani, Gertler y Schargrodsky (2005), sobre la política de aguas en Argentina; Ferraz y Finan (2008) sobre auditorías y

desempeño de los políticos, y Chetty, Friedman y Sáez (2013) sobre los créditos fiscales en Estados Unidos. En el recuadro 16.1 se presenta un ejemplo de evaluación de impacto de salud en Argentina. Por su parte, el recuadro 16.2 ilustra el uso de datos administrativos en la evaluación de impacto de un programa de transferencias monetarias en Honduras.

En algunos casos, los datos requeridos para la evaluación de impacto se pueden recopilar implementando nuevos sistemas de información o de datos administrativos. Esta implementación se puede coordinar con la de un diseño de evaluación, de modo que los indicadores de resultado se recopilen para un grupo de tratamiento y un grupo de comparación en múltiples

Recuadro 16.1: Elaboración de una base de datos en la evaluación del Plan Nacer de Argentina

Al evaluar el programa de financiamiento de la salud basado en resultados de Argentina, Plan Nacer, Gertler, Giovagnoli y Martínez (2014) combinaron datos administrativos de diversas fuentes para formar una base de datos grande y exhaustiva para el análisis. Después de la falta de éxito de diversas estrategias de evaluación anteriores, los investigadores adoptaron un enfoque de variables instrumentales. Esto requirió una cantidad sustancial de datos del universo de los registros de nacimientos de las siete provincias estudiadas.

Los investigadores necesitaban datos sobre la atención prenatal y los resultados al nacer, que se podían encontrar en los registros de nacimientos de los hospitales públicos. Luego tuvieron que determinar si la madre era beneficiaria del Plan Nacer y si la clínica que visitaba estaba incorporada en el programa en el momento de la visita. Para construir una base de datos con toda esta información, el equipo de evaluación vinculó cinco diferentes fuentes de datos, incluyendo las bases de datos de los hospitales públicos de maternidad, los datos de implementación del programa Plan Nacer, los registros farmacéuticos, el censo de población de 2001 e información geográfica de los centros de salud. La obtención de historiales médicos de los nacimientos individuales en los hospitales de maternidad fue una de las tareas más difíciles. Cada hospital de maternidad recopilaba datos sobre la atención prenatal y los resultados al nacer, pero solo cerca de la mitad de los registros estaban digitalizados. El resto se componía de documentos en papel, por lo que el equipo de evaluación tuvo que ingresar los registros en papel en el sistema computarizado.

El equipo compiló una base de datos exhaustiva del 78% de los nacimientos ocurridos durante el período de evaluación. Esto generó una base de datos grande que les permitió examinar el impacto del Plan Nacer en sucesos relativamente raros, como la mortalidad neonatal. Normalmente, esto no es posible en las evaluaciones con muestras más pequeñas recopiladas a través de encuestas. La evaluación determinó que los beneficiarios del Plan Nacer tenían una probabilidad un 74% menor de mortalidad neonatal en el hospital que los no beneficiarios.

Fuente: Gertler, Giovagnoli y Martínez (2014).

Recuadro 16.2: Utilización de datos censales para reevaluar el PRAF en Honduras

El objetivo del Programa de Asignación Familiar (PRAF), de Honduras, es mejorar los resultados educativos y de salud de los niños pequeños que viven en condiciones de pobreza. Proporciona transferencias monetarias a hogares elegibles en función de la asistencia escolar y de las visitas a un centro de salud de manera regular. El programa comenzó en 1990. Un componente de la evaluación se incluyó en la segunda fase del PRAF en 1998. Glewwe y Olinto (2004) y Morris et al. (2004) informaron de impactos positivos en los resultados en educación y salud.

Varios años más tarde, Galiani y McEwan (2013) reevaluaron el impacto del programa, utilizando una fuente de datos diferente. Mientras que la evaluación de impacto original había recopilado datos de encuesta de 70 municipios sobre un total de 298, Galiani y McEwan utilizaron datos del censo de Honduras de 2001. Así, fusionaron los datos individuales y de los hogares del censo con los datos a nivel municipal sobre las comunidades tratadas. Esto proporcionó a los investigadores un tamaño muestral más grande, lo que les permitió poner a prueba la robustez de los hallazgos, además de los efectos de derrame. Por otro lado, dado que contaban con datos del censo de todos los municipios, pudieron aplicar dos diseños de regresión discontinua diferentes utilizando grupos de comparación alternados. Para el primer diseño de regresión discontinua, utilizaron el umbral de elegibilidad; para el segundo, los límites del municipio.

Al igual que en las evaluaciones de impacto anteriores, Galiani y McEwan encontraron impactos positivos y estadísticamente significativos a partir del programa. Sin embargo, sus estimaciones indicaban que el PRAF había tenido un impacto mucho mayor que el impacto encontrado en la evaluación original. Observaron que el PRAF aumentaba la matriculación escolar en los niños elegibles en un 12% más que en el grupo de comparación. Los resultados de los diseños de regresión discontinua alternados generalmente confirman la robustez de las conclusiones.

Fuente: Galiani y McEwan (2013).

ocasiones. Puede que la puesta en marcha de sistemas de información se produzca antes de lanzar las nuevas intervenciones, de modo que los centros administrativos en el grupo de comparación utilicen el nuevo sistema de información antes de recibir la intervención que se evalúa. Dado que la calidad de los datos administrativos puede variar, requieren una auditoría y una verificación externa para garantizar la fiabilidad de la evaluación. Recoger datos de la evaluación de impacto a través de fuentes administrativas en lugar de hacerlo a través de encuestas puede reducir drásticamente el costo de una evaluación, pero no siempre es viable.

Aun cuando los datos existentes no sean suficientes para toda la evaluación de impacto, a veces pueden ser utilizados en partes de la evaluación.

Por ejemplo, en algunos casos, los programas recopilan datos detallados sobre beneficiarios potenciales para definir quién es elegible. O puede que los datos del censo estén disponibles poco antes de que un programa se implemente. En esos casos, los datos existentes a veces pueden ser utilizados para documentar un equilibrio de línea de base, en las características anteriores al programa, entre los grupos de tratamiento y comparación, aunque se seguirían necesitando datos de seguimiento adicionales para medir un conjunto más amplio de indicadores de resultados.

La recopilación de datos de nuevas encuestas

Los datos existentes son suficientes para toda una evaluación de impacto solo en casos relativamente raros. Si los datos administrativos no son suficientes para la evaluación, muy posiblemente habrá que depender de datos de encuestas. Como consecuencia, lo más probable es que se tenga que presupuestar la recopilación de nuevos datos. A pesar de que la recopilación de datos suele implicar el mayor costo de una evaluación de impacto, también puede ser una inversión de alto retorno de la que a menudo depende la calidad de la evaluación. La recopilación de nuevos datos proporciona la flexibilidad para garantizar que se midan todos los indicadores necesarios para una evaluación integral del desempeño del programa.

La mayoría de las evaluaciones de impacto requieren recopilar datos de encuestas, incluyendo al menos una *encuesta de línea de base* antes de la intervención o innovación que se evalúa, y una *encuesta de seguimiento* después de que se ha implementado la intervención. Los datos de las encuestas pueden ser de diversos tipos, en función del programa que se evalúa y de la unidad de análisis. Por ejemplo, las encuestas de empresas utilizan a las firmas como la principal unidad de observación, las encuestas de instalaciones utilizan los centros de salud o las escuelas como la principal unidad de observación, y las encuestas de hogares utilizan los hogares como la principal unidad de observación. La mayoría de las evaluaciones dependen de encuestas individuales o de hogares como fuente primaria de datos. En esta sección, se revisan algunos principios generales de la recopilación de datos de las encuestas. Aunque estos se refieren sobre todo a las encuestas de hogares, los mismos principios son válidos para la mayoría de otros tipos de datos de encuestas.

El primer paso para decidir si utilizar los datos existentes o recopilar nuevos datos mediante encuestas será determinar el enfoque del muestreo, así como el tamaño necesario de la muestra (como se analizó en el capítulo 15). Una vez que se decida recopilar datos de encuestas para la evaluación, habrá que:

- Determinar quién recopilará los datos.
- Desarrollar y poner a prueba el instrumento de recopilación de datos.
- Llevar a cabo un trabajo de campo y realizar el control de calidad.
- Procesar y almacenar los datos.

La implementación de estos diversos pasos se suele contratar, pero es esencial que se comprendan su alcance y sus componentes clave para poder gestionar efectivamente una evaluación de impacto de calidad.

Determinar quién recopilará los datos

Es necesario designar con antelación a la agencia encargada de recopilar los datos. Al momento de decidir quién lo haría, habrá varias consideraciones. Los candidatos potenciales para esta tarea son:

- La institución a cargo de implementar el programa.
- Otra institución gubernamental con experiencia en la recopilación de datos (como una agencia estadística nacional).
- Una empresa independiente o institución especializada en recopilación de datos.

La entidad que recopile los datos siempre tiene que coordinarse estrechamente con el organismo que implemente el programa. Se requiere una estrecha coordinación para garantizar que las operaciones del programa no se pongan en marcha antes de recopilar los datos de línea de base. Cuando se necesitan datos de línea de base para el funcionamiento del programa (por ejemplo, datos para un índice de focalización, en el contexto de una evaluación basada en un diseño de regresión discontinua), la entidad encargada de la recopilación de datos debe ser capaz de procesar los mismos rápidamente y transmitirlos a la institución encargada de las operaciones del programa. También se requiere una coordinación estrecha de la programación de la recopilación de datos de la encuesta de seguimiento. Por ejemplo, si se ha elegido una implementación con asignación aleatoria, la encuesta de seguimiento debe llevarse a cabo antes de que el programa se ponga en marcha en el grupo de comparación, para evitar la contaminación.

Un factor sumamente importante en la decisión de quién recopilará los datos es la utilización de los mismos procedimientos de recopilación de datos en los grupos de comparación y de tratamiento. A menudo la agencia de implementación tiene contacto solo con el grupo de tratamiento y no está en una buena posición para recopilar datos de los grupos de comparación. Sin embargo, utilizar diferentes organismos de recopilación de datos para

los grupos de tratamiento y de comparación es muy riesgoso, dado que esto puede crear diferencias en los resultados medidos en los dos grupos sencillamente porque los procedimientos de recopilación de datos son diferentes. Si la agencia ejecutora no puede recopilar datos efectivamente de los grupos de tratamiento y de comparación, debería contemplarse seriamente la posibilidad de contratar una institución o agencia externa.

Concepto clave
Deben utilizarse los mismos procedimientos de recopilación de datos tanto en los grupos de comparación como de tratamiento.

En algunos contextos, también puede ser recomendable contratar una agencia independiente para recoger los datos con el fin de garantizar que estos se consideren objetivos. Puede que no se justifiquen las preocupaciones de que la agencia ejecutora del programa no recopile datos objetivos, pero un organismo de recopilación de datos independiente que no tiene intereses en juego en los resultados de la evaluación añade credibilidad al esfuerzo general de evaluación de impacto. También puede garantizar que los encuestados no perciban la encuesta como parte del programa y, de esta manera, se minimiza el riesgo de que los encuestados den respuestas estratégicas intentando aumentar lo que perciben como la posibilidad de participar en un programa.

Dado que la recopilación de datos comprende una secuencia compleja de operaciones, se recomienda que una entidad especializada y experimentada sea la responsable. Hay pocos organismos ejecutores de programas con suficiente experiencia para recopilar los datos a gran escala y preservar la calidad, ambos criterios necesarios para una evaluación de impacto. En la mayoría de los casos, se tendrá que pensar en contratar a una institución local, como una agencia estadística nacional o una empresa o *think tank* especializados.

La contratación de una institución local, como una agencia estadística nacional, puede exponer a la institución a los estudios de evaluación de impacto y contribuir a mejorar su capacidad, lo cual en sí mismo puede ser un beneficio secundario de la evaluación de impacto. Sin embargo, las agencias estadísticas nacionales no siempre tendrán la capacidad logística para asumir otros encargos además de sus actividades regulares. Puede que también carezcan de la experiencia necesaria para llevar a cabo encuestas para las evaluaciones de impacto, como la experiencia de efectuar un seguimiento exitoso de los individuos a lo largo del tiempo, o para aplicar instrumentos de encuesta no tradicionales. Si estas limitaciones aparecen, contratar una empresa independiente o institución especializada en recopilación de datos puede ser lo más práctico.

No siempre es necesario que sea la misma entidad la que recopila información de las encuestas de línea de base y de seguimiento, dado que pueden variar en su alcance. Por ejemplo, en una evaluación de impacto de un programa de capacitación cuya población de interés está compuesta por los individuos que se inscribieron en el curso, la institución a cargo del curso podría recopilar los datos de línea de base cuando los individuos se inscriben.

Sin embargo, es poco probable que la misma agencia también sea la mejor opción para recopilar información de seguimiento, tanto para los grupos de tratamiento como de comparación. En este contexto, contratar rondas de recopilación de datos por separado tiene sus ventajas, pero se debería hacer un esfuerzo para no perder información entre las rondas, información que será útil para hacer un seguimiento de los hogares o de los individuos, así como para asegurar que los datos de línea de base y de seguimiento se midan de manera consistente.

A fin de decidir cuál es la mejor institución para recopilar los datos de la evaluación de impacto, deben sopesarse todos estos factores (experiencia en la recopilación de datos, capacidad de coordinar con la agencia ejecutora del programa, independencia, oportunidades para la mejora de capacidades, adaptabilidad al contexto de la evaluación de impacto), junto con el costo previsto y la probable calidad de los datos obtenidos en cada caso. Una manera efectiva de identificar la organización mejor situada para recopilar datos de calidad consiste en redactar términos de referencia claros y pedir a las organizaciones que presenten propuestas técnicas y financieras.

Dado que la entrega oportuna y la calidad de los datos suelen ser cruciales para la fiabilidad de la evaluación de impacto, el contrato para la agencia encargada de la recopilación de datos debe estructurarse con gran cuidado. El alcance del trabajo previsto y los productos deben definirse con suma claridad. Además, se recomienda introducir incentivos en los contratos y vincular esos incentivos a indicadores claros de la calidad de los datos. Por ejemplo, la tasa de falta de respuesta es un indicador clave de la calidad de los datos. Para crear incentivos con el fin de que las agencias de recopilación de datos minimicen las no respuestas, el contrato puede estipular un costo unitario para el primer 80% de la muestra, un costo unitario superior para las unidades de entre el 80% y el 90% y, una vez más, un costo unitario superior para las unidades de entre el 90% y el 100%. Como alternativa, se puede redactar un contrato por separado para que la empresa encuestadora realice un seguimiento de los no encuestados. Además, el contrato de la empresa de recopilación de datos puede incluir incentivos o condiciones relacionadas con la verificación de la calidad de los datos, como comprobaciones externas o auditorías de calidad de una submuestra de la encuesta de evaluación de impacto.

La elaboración del instrumento de recopilación de datos y las pruebas piloto

Al contratar la recopilación de datos, el equipo de evaluación desempeña un rol clave proporcionando orientación específica sobre el contenido de los instrumentos o cuestionarios de la recopilación de datos. Los instrumentos de recopilación de datos deben obtener toda la información requerida para

responder a la pregunta de las políticas definida por la evaluación de impacto. Como ya se ha señalado, los *indicadores* deben medirse a lo largo de la cadena de resultados, lo que incluye los indicadores de los resultados finales, los resultados intermedios y las medidas de los beneficios y la calidad de la implementación del programa.

Es importante ser selectivo acerca de qué indicadores medir. Ser selectivo contribuye a limitar los costos de recopilación de datos, simplifica la tarea de la agencia de recopilación y mejora la calidad de los datos recopilados minimizando las demandas de tiempo para encuestadores y encuestados. La recopilación de información que es irrelevante o que probablemente no se utilizará tiene un costo muy alto. Los datos adicionales requieren más tiempo de preparación, capacitación, recopilación y procesamiento. Con una disponibilidad y una capacidad de atención limitadas, puede que los encuestados proporcionen información de calidad cada vez más inferior a medida que la encuesta avanza, y los entrevistadores tendrán incentivos extra para ahorrar tiempo con el fin de cumplir con sus objetivos de la encuesta. Por lo tanto, las preguntas superfluas no son "gratis". Tener objetivos claros para la evaluación de impacto, alineados con objetivos del programa bien definidos, puede ayudar a priorizar la información necesaria. Un plan de preanálisis elaborado con antelación (véanse los detalles en los capítulos 12 y 13) contribuirá a asegurar que la encuesta recopile los datos requeridos para el análisis de impacto y evitar la inclusión de información superflua (y costosa).

Es preferible recopilar datos sobre los indicadores de resultado y las características de control de manera consistente en la línea de base y en el seguimiento. Contar con datos de línea de base es sumamente recomendable. Aun cuando se utilice una asignación aleatoria o un diseño de regresión discontinua, donde en principio se pueden usar sencillas diferencias después de la intervención para estimar el impacto de un programa, los datos de línea de base son esenciales para probar si el diseño de la evaluación de impacto es adecuado (véase el debate en la segunda parte). Contar con datos de línea de base puede servir como póliza de seguro cuando la asignación aleatoria no funciona, en cuyo caso se pueden utilizar métodos de diferencias en diferencias como alternativa. Los datos de línea de base también son útiles durante la etapa de análisis del impacto, dado que las variables de control de línea de base pueden contribuir a aumentar la potencia estadística y permitir analizar impactos en diferentes subpoblaciones. Por último, los datos de línea de base pueden utilizarse para mejorar el diseño del programa. Por ejemplo, los datos de línea de base a veces permiten analizar la eficiencia focalizada o proporcionan información adicional sobre los beneficiarios a la agencia que implementa el programa. En algunos casos, la encuesta de seguimiento puede incluir un conjunto más amplio de indicadores que la encuesta de línea de base.

Una vez que se han definido los datos centrales que se debe recopilar, el próximo paso consiste en determinar exactamente cómo medir esos indicadores. La *medición* es un arte en sí misma y es preferible que de ella se ocupen los especialistas, entre los cuales se hallan el equipo de investigación de la evaluación de impacto, la agencia contratada para recopilar datos, los expertos de las encuestas y los expertos en la medición de indicadores complejos específicos. Los indicadores de resultado deberían ser lo más consistentes posible con las mejores prácticas locales e internacionales. Siempre es útil tener en cuenta cómo los indicadores de interés han sido medidos en encuestas similares, tanto a nivel local como internacional. Utilizar los mismos indicadores (lo que incluye los mismos módulos o preguntas de las encuestas) garantiza la comparabilidad entre los datos preexistentes y los datos recopilados para la evaluación de impacto. Elegir un indicador que no sea plenamente comparable o no esté bien medido puede limitar la utilidad de los resultados de la evaluación. En algunos casos, puede que tenga sentido invertir los recursos necesarios para recopilar el nuevo indicador de resultado "innovador", así como una alternativa más establecida.

Concepto clave

La medición de los indicadores es un arte y es necesario que sea gestionada por los especialistas, entre ellos: el equipo de investigación de la evaluación de impacto, la agencia contratada para recopilar datos, los expertos de encuestas y los expertos en la medición de indicadores específicos complejos.

Se debe prestar particular atención para asegurar que todos los indicadores se puedan medir exactamente de la misma manera para todas las unidades tanto del grupo de tratamiento como de comparación. La utilización de diferentes métodos de recopilación de datos (por ejemplo, una encuesta telefónica para un grupo y una encuesta presencial para otro) crea el riesgo de generar sesgos. Lo mismo sucede con la recopilación de datos en diferentes momentos para los dos grupos (por ejemplo, recopilar datos para el grupo de tratamiento durante la estación de lluvias y para el grupo de comparación durante la estación seca). Por esto, los procedimientos empleados para medir cualquier indicador de resultado deberían formularse con suma precisión. El proceso de recopilación de datos debe ser exactamente el mismo para todas las unidades. En un cuestionario, cada módulo relacionado con el programa debe introducirse sin afectar el flujo o la formulación de las respuestas en otras partes del cuestionario. De hecho, cuando sea posible, es preferible evitar hacer cualquier distinción entre los grupos de tratamiento y comparación en el proceso de recopilación de datos. En la mayoría de los casos, la agencia que lleve a cabo la recopilación (o al menos los encuestadores individuales) no debe tener motivos para conocer la condición de tratamiento o de comparación de los individuos en la encuesta.

Una decisión crucial que hay que tomar se relaciona con la forma de medir los indicadores de resultados, es decir: si se hace mediante encuestas tradicionales basadas en cuestionarios y preguntas auto-reportadas o a través de otros métodos. En los últimos años, se han producido varios avances para medir resultados o conductas clave que son relevantes en las evaluaciones de impacto. Los avances incluyen el perfeccionamiento de los métodos

para recopilar datos auto-reportados mediante cuestionarios, así como técnicas para medir directamente los resultados clave.

El *diseño del cuestionario* ha sido objeto de importantes investigaciones. Se han escrito libros enteros sobre la mejor manera de medir indicadores particulares en contextos específicos, incluyendo cómo redactar las preguntas formuladas en las encuestas de hogares.[1] También hay una base de evidencia creciente sobre la mejor manera de diseñar cuestionarios para recopilar datos agrícolas, datos sobre el consumo o datos de empleo para maximizar su precisión.[2] Parte de la evidencia reciente proviene de experimentos aleatorios que prueban diferentes maneras de estructurar cuestionarios y comparar su fiabilidad.[3] De la misma manera, el diseño del cuestionario requiere prestar atención a las mejores prácticas internacionales y a las experiencias locales en materia de medición de indicadores. Pequeños cambios en la redacción o en la secuencia de las preguntas pueden tener efectos sustanciales en los datos recopilados, de modo que en el desarrollo del cuestionario es esencial prestar atención a los detalles. Esto es especialmente importante cuando se intenta asegurar la comparabilidad entre diferentes encuestas, lo que implica, por ejemplo, medir los resultados repetidas veces a lo largo del tiempo. El recuadro 16.3 aborda las directrices relacionadas con el diseño del cuestionario y proporciona otras referencias.

Recuadro 16.3: Diseño y formato de los cuestionarios

Aunque el diseño del cuestionario en las evaluaciones de impacto es una parte integral de la calidad de los datos, a menudo se lo ignora. El diseño de un cuestionario es un proceso complejo, extenso e iterativo que comprende numerosas decisiones a lo largo del camino a propósito de lo que se puede medir y cómo medirlo. El curso de métodos de evaluación de impacto aplicada de la Universidad de California, Berkeley (http://aie.cega.org) proporciona una guía para el diseño del cuestionario, en el cual destacan tres fases: contenido, redacción y puesta a prueba. A lo largo de estas fases, el módulo subraya la importancia de involucrar a las partes interesadas pertinentes, y dedicar tiempo suficiente a las iteraciones repetidas y a pruebas rigurosas:

1. *Contenido.* Se determina el contenido de una encuesta empezando por definir los efectos que hay que medir, las unidades de observación y las correlaciones con otros factores. Estas definiciones conceptuales luego tendrán que ser traducidas en indicadores concretos.

2. *Redacción.* Se redactan las preguntas para medir los indicadores seleccionados. Se trata de un paso crucial, dado que la calidad de los datos depende de ello.

Continúa en la página siguiente.

Recuadro 16.3: Diseño y formato de los cuestionarios *(continúa)*

El módulo ofrece recomendaciones más detalladas sobre la redacción de las preguntas, la organización de la encuesta, el formato y otras consideraciones clave.

3. *Pruebas*. El cuestionario es probado en tres niveles: la pregunta, el módulo y el conjunto de la encuesta.

El formato del cuestionario también es importante para asegurar datos de calidad. Dado que diferentes maneras de formular la misma pregunta en la encuesta puede generar respuestas diferentes, tanto el marco como el formato de las preguntas debería ser el mismo para todas las unidades con el fin de evitar sesgos de los encuestados o los encuestadores. Naciones Unidas (2005) formula seis recomendaciones específicas en relación con el formato de los cuestionarios en las encuestas de hogares. Estas recomendaciones se aplican también a la mayoría de los demás instrumentos de recopilación de datos:

1. Se debe redactar cada pregunta detalladamente en el cuestionario de manera que el entrevistador pueda dirigir la entrevista leyendo cada pregunta palabra por palabra.
2. El cuestionario debe incluir definiciones precisas de todos los conceptos clave utilizados en la encuesta, de modo que el entrevistador se pueda referir a la definición durante la entrevista si fuera necesario.
3. Las preguntas deben ser lo más breves y sencillas posible, y deben utilizar un lenguaje común y comprensible.
4. Los cuestionarios deben diseñarse de manera que las respuestas a casi todas las preguntas estén precodificadas.
5. El esquema de la codificación de las respuestas debería ser consistente en todas las preguntas.
6. La encuesta debería incluir patrones de salto, que indican qué preguntas no deberían formularse, sobre la base de las respuestas dadas a las preguntas anteriores.

Una vez que la persona encargada para trabajar en el instrumento ha redactado un cuestionario, este debe presentarse a un equipo de expertos para ser debatido. Se debe consultar a todos los que participan en el equipo de evaluación (responsables de las políticas, investigadores, analistas de datos y recopiladores de datos) a propósito de si el cuestionario recoge toda la información que se desea, de manera adecuada. La revisión de un equipo de expertos es necesaria pero no suficiente, dado que la puesta a prueba intensiva en el terreno siempre es primordial.

Se ha ido desarrollado cada vez más un conjunto de técnicas para obtener la *medición directa de resultados*. Por ejemplo, en el sector de la salud, a veces se utilizan casos clínicos para presentar síntomas concretos a los trabajadores de la salud y evaluar si el proveedor recomienda el tratamiento adecuado sobre la base de directrices y protocolos establecidos. Estos casos clínicos proporcionan una medida directa de los

conocimientos de los proveedores de la salud. Las evaluaciones recientes cuentan con pacientes estandarizados (también conocidos como pacientes de incógnito o simulados) que visitan los centros de salud y evalúan directamente la calidad de los servicios prestados.[4] En el sector de educación, numerosas evaluaciones pretenden calcular los impactos de los programas en el aprendizaje de los alumnos. Para ello, se utiliza una gama de evaluaciones del aprendizaje o mediciones directas de las habilidades de los alumnos. También se han desarrollado varias baterías de pruebas para medir directamente el desarrollo cognitivo, lingüístico o motor de los niños pequeños en el contexto de las evaluaciones de impacto de las intervenciones de desarrollo infantil temprano (DIT). Asimismo, se ha progresado en la obtención de mediciones directas de las habilidades de los adultos, incluidas las habilidades socioemocionales o los rasgos de la personalidad. Además de la medición directa de las habilidades, un número creciente de evaluaciones de impacto apunta a obtener mediciones de la calidad de la enseñanza a través de las observaciones directas de la conducta de los profesores en el aula.

La observación directa de resultados clave es particularmente importante cuando se hace difícil obtener verazmente los resultados de interés de parte de los encuestados. Por ejemplo, para evitar depender de los datos auto-reportados para medir los resultados relacionados con los delitos o la violencia, algunas evaluaciones de impacto han incorporado investigadores capacitados en las comunidades de la muestra para que observen directamente la conducta de los sujetos con métodos etnográficos. Esta observación directa puede eludir los problemas relacionados con conductas auto-reportadas y proporcionar una información más precisa cuando se lleva a cabo adecuadamente. Los recientes avances tecnológicos también permiten mediciones directas de toda una gama de conductas humanas y, por lo tanto, pueden contribuir a limitar el uso de datos auto-reportados. Entre otros ejemplos, cabe señalar la observación directa de la programación en el tiempo y la intensidad en el uso de cocinas mejoradas, y las mediciones directas de la calidad del agua, del uso de letrinas y de la temperatura interior utilizando sensores electrónicos.

Las evaluaciones de impacto suelen depender de una mezcla de encuestas basadas en cuestionarios tradicionales y otros métodos con los que observar directamente los resultados de interés. Por ejemplo, en el contexto de la evaluación de impacto del financiamiento basado en resultados en el sector de la salud, se mide una gama de indicadores a través de fuentes complementarias (Vermeersch, Rothenbühler y Sturdy, 2012). Una encuesta de un centro de salud comprende una evaluación del centro para medir sus principales características, una entrevista con un trabajador de la salud para estimar las características de este, y entrevistas de salida con los pacientes

para valorar los servicios prestados, así como indicadores de la calidad de los cuidados mediante una mezcla de casos clínicos y observación directa. Las encuestas de hogares incluyen datos a nivel de los hogares sobre la conducta de estos y de los individuos, como la frecuencia de las visitas al centro, los cuidados recibidos y los gastos en salud, así como también módulos a nivel individual de la salud de las mujeres y de los niños. Además de mediciones antropométricas, se realizan pruebas biomédicas para medir directamente la prevalencia de la anemia, la malaria o el VIH. Por último, los cuestionarios de las comunidades capturan características de la comunidad, los servicios, la infraestructura, el acceso a los mercados, los precios y los shocks a nivel comunitario.

Además de desarrollar indicadores y encontrar la manera más adecuada para medirlos, otra decisión clave al recopilar nuevos datos es la tecnología de recopilación utilizada. Los métodos tradicionales recopilan los datos en papel y luego los digitalizan, a menudo mediante un enfoque de entrada de datos de doble ciego, que implica la presencia de dos agentes que digitan la misma información por separado antes de comparar los datos para verificar las imprecisiones. Con los recientes avances tecnológicos, los instrumentos de recopilación de datos asistidos por computador se han vuelto prevalentes. La recopilación de datos mediante aplicaciones instaladas en teléfonos inteligentes o *Tablets* puede acelerar el procesamiento de datos, y al mismo tiempo proporciona oportunidades para realizar verificaciones de la calidad de los datos y validarlos en tiempo real. En el recuadro 16.4 se abordan algunas de las ventajas y desventajas de la recopilación electrónica de datos.

Es sumamente importante que el instrumento de recopilación de datos sea probado en el terreno extensamente antes de finalizar. La realización de *pruebas piloto* amplias del instrumento pondrá a prueba su adecuación al contexto local y su contenido, y cualquier formato y opciones de redacción alternativas, así como también los protocolos de recopilación de datos, incluida la tecnología. Probar el instrumento de recopilación de datos en *pruebas en el terreno* es crucial para chequear su duración y para verificar que su formato sea suficientemente consistente y exhaustivo para producir mediciones precisas de toda la información relevante. Las pruebas en el terreno constituyen una parte integral de la elaboración de los instrumentos de recopilación de datos.

Dirección del trabajo de campo y gestión del control de calidad

Aun cuando se contrate la recopilación de datos con una entidad externa, es crucial tener una comprensión clara de todos los pasos involucrados en este proceso para garantizar que se hayan establecido los *mecanismos de control*

Recuadro 16.4: Algunas ventajas y desventajas de la recopilación electrónica de datos

Las entrevistas personales asistidas por computador (CAPI, por sus siglas en inglés, *computer-assisted personal interviewing*) brindan una alternativa a las entrevistas tradicionales de lápiz y papel (PAPI, *pen-and-paper interviewing*). En las CAPI, primero se descarga la encuesta en un aparato electrónico, como una *Tablet* o un teléfono inteligente. El entrevistador lee las preguntas en la pantalla e ingresa inmediatamente las respuestas en el programa. Se han desarrollado diversos programas y aplicaciones para la recopilación de datos en CAPI. No obstante, el equipo de evaluación debe considerar detenidamente los pros y contras de las CAPI.

Algunas ventajas:

- La recopilación electrónica de datos puede mejorar la calidad de los datos. En un experimento aleatorio diseñado para comparar CAPI y PAPI para una encuesta sobre consumo en Tanzania, Caeyers, Chalmers y De Weerdt (2012) observaron que los datos de las encuestas en papel contenían errores que se evitaron en las encuestas electrónicas. Los investigadores descubrieron que los errores en los datos del PAPI estaban correlacionados con ciertas características de los hogares, lo cual puede crear sesgos en algunos análisis de datos.
- Los programas de recopilación electrónica de datos pueden incluir sistemas de verificación de consistencia automatizados. Ciertas respuestas pueden activar mensajes de alerta de manera que los errores de ingreso de los datos se minimizan y cualquier problema se aclara con el encuestado durante la entrevista. Por ejemplo, Fafchamps et al. (2012) estudiaron los beneficios del control de la consistencia en una encuesta de microempresas en Ghana. Observaron que cuando se introducían los controles de consistencia, la desviación estándar de los datos sobre beneficios y ventas era menor. Sin embargo, también observaron que la mayor parte del tiempo no se requería una corrección: entre el 85% y el 97% de las veces, los encuestados confirmaban la respuesta original.
- Las entrevistas pueden ser más breves y más fáciles. Cuando se utilizan las CAPI, el flujo del cuestionario se puede personalizar para orientar mejor a los entrevistadores a través de patrones de salto, y minimizar los errores y omisiones en el cuestionario. En una encuesta de hogares conducida en Tanzania, las entrevistas CAPI fueron, en promedio, un 10% más breves que los cuestionarios similares recopilados en papel, según las observaciones de Caeyers, Chalmers y De Weerdt (2012).
- La recopilación electrónica de datos elimina la necesidad del reingreso manual de los datos. Esto puede reducir costos y acelerar el procesamiento.
- El uso de la tecnología puede aportar una gama de beneficios indirectos. Por ejemplo, al utilizar *Tablets* o teléfonos inteligentes, es fácil registrar las coordenadas de GPS o tomar fotos. También se pueden introducir variaciones experimentales en el contenido de la encuesta. Con algunos programas, ciertas partes de la entrevista se pueden grabar con el fin de facilitar la calidad y los controles del monitoreo.

Continúa en la página siguiente.

Recuadro 16.4: Algunas ventajas y desventajas de la recopilación electrónica de datos *(continúa)*

Algunas desventajas:

- Los costos fijos tienden ser más elevados en las CAPI que en las PAPI, aunque los costos variables pueden ser más bajos. El costo inicial de comprar y programar los aparatos electrónicos puede resultar demasiado alto para los presupuestos más pequeños de evaluación de impacto. También se necesita más tiempo al comienzo para asegurar una programación y pruebas adecuadas de los cuestionarios electrónicos, que a menudo se producen después de que ya se han elaborado los cuestionarios en papel.
- Se requieren conocimientos técnicos expertos específicos para programar los cuestionarios electrónicos y crear procesos para gestionar el flujo de datos recopilados electrónicamente. En los países en desarrollo con baja capacidad en materia de tecnologías de la información, a veces esto se torna difícil de conseguir. También es más arduo desarrollar programas para cuestionarios que no estén en inglés o en una lengua romance.
- Los problemas tecnológicos pueden perturbar la recopilación de datos o dañar la consolidación de datos en un sitio seguro. Pueden surgir problemas durante la recopilación de datos, cuando el aparato electrónico tiene una pantalla pequeña o una interfaz con la que los entrevistadores no están familiarizados. El riesgo de robo también es mayor en el caso de los aparatos electrónicos en comparación con las encuestas en papel. Por último, la consolidación y sincronización de los datos en un sitio seguro requiere protocolos claros para minimizar el riesgo de pérdida de datos. La transmisión electrónica de datos es conveniente pero requiere un nivel mínimo de conectividad.

Fuente: Caeters, Chalmers y De Weerdt (2012); Fafchamps et al. (2012).

de calidad requeridos y los *incentivos correctos*. La entidad encargada de recopilar los datos tendrá que coordinar el trabajo de un gran número de actores diferentes, entre ellos los encuestadores, supervisores, coordinadores en el terreno y personal de apoyo logístico, además del equipo que ingresa los datos, compuesto por los programadores, los supervisores y los operadores del ingreso de datos. Debe establecerse un *plan de trabajo* claro para coordinar la labor de todos estos equipos y ese plan de trabajo es un elemento clave.

Antes de que comience la recopilación de datos, el plan de trabajo debe incluir una *capacitación* adecuada del equipo de recopilación de datos. Se debe elaborar un *manual de referencia* completo para la capacitación, el cual debe utilizarse a lo largo del trabajo de campo. La capacitación es clave para asegurar que todos los que participan recopilen los datos de manera consistente. El proceso de capacitación también es una buena oportunidad para identificar a los mejores encuestadores y para llevar a cabo una última prueba

piloto de los instrumentos y los procedimientos en condiciones normales. Una vez que se haya elaborado la muestra, que los instrumentos hayan sido diseñados y probados en pruebas piloto, y que los equipos hayan sido capacitados, puede comenzar la recopilación de datos. Es una buena práctica asegurar que el plan de trabajo de campo que tiene cada equipo de la encuesta recopile datos sobre el mismo número de unidades de tratamiento y comparación.

Como se señaló en el capítulo 15, el muestreo adecuado es esencial para asegurar la calidad de la muestra. Sin embargo, mientras se recopilan los datos pueden producirse numerosos *errores de no muestreo*. En el contexto de una evaluación de impacto, una preocupación particular es que aquellos errores pueden no ser los mismos en los grupos de tratamiento y de comparación.

Concepto clave

La no respuesta surge cuando faltan datos o hay datos incompletos para algunas unidades de la muestra. La no respuesta puede crear sesgos en los resultados de la evaluación.

La *falta de respuesta* surge cuando se vuelve imposible recopilar todos los datos para algunas unidades de la muestra. Dado que las muestras reales utilizadas para el análisis se limitan a aquellas unidades para las que se pueden recopilar datos, las unidades que deciden no responder a una encuesta pueden volver la muestra menos representativa y crear un sesgo en los resultados de la evaluación. El *desgaste de la muestra* es una forma habitual de no respuesta que se produce cuando algunas unidades abandonan la muestra entre las rondas de recopilación de datos; por ejemplo, los migrantes, de los que es difícil hacer un seguimiento.

El desgaste de la muestra debido a la no respuesta es especialmente problemático en el contexto de las evaluaciones de impacto porque puede crear diferencias entre el grupo de tratamiento y el grupo de comparación. Por ejemplo, el *desgaste de la muestra* puede ser diferente en los dos grupos: si los datos se recopilan después de que el programa ha comenzado a implementarse, la tasa de respuesta entre las unidades de tratamiento puede ser más elevada que entre las unidades de comparación. Esto puede ocurrir porque estas últimas se muestran descontentas por no haber sido seleccionadas o porque es más probable que migren. Las no respuestas también se pueden producir en el propio cuestionario, normalmente porque faltan algunos indicadores o porque los datos para una unidad particular son incompletos.

Los errores de medición constituyen otro tipo de problema, que puede generar sesgos si tiene lugar de forma sistemática. El *error de medición* es la diferencia entre el valor de una característica tal como la presenta el encuestado y el valor verdadero (pero desconocido) (Kasprzyk, 2005). Esta diferencia se explica por la manera en que el cuestionario está redactado o por el método de recopilación de datos elegido, o puede producirse debido a los entrevistadores que están llevando a cabo la encuesta o al encuestado que responde.

La calidad de la evaluación de impacto depende directamente de la calidad de los datos recopilados. Es necesario especificar los *estándares de calidad* para todas las partes interesadas en el proceso de recopilación de datos; estos estándares deberían subrayarse particularmente durante la

capacitación de los encuestadores y en los manuales de referencia. Por ejemplo, es esencial contar con procedimientos detallados para minimizar la no respuesta o (si es aceptable) reemplazar unidades en la muestra. La agencia de recopilación de datos debe entender claramente las tasas aceptables de no respuesta y de desgaste de la muestra. Como referencia, numerosas evaluaciones de impacto se proponen mantener la no respuesta y el desgaste por debajo del 5%. El objetivo dependerá de la programación en el tiempo de la evaluación de impacto y de la unidad de análisis: se esperaría que el desgaste fuera menor en una encuesta que se produce poco después de la encuesta de línea de base, y relativamente más alto para las evaluaciones de impacto de largo plazo que siguen a los individuos muchos años más tarde. También se esperarían tasas de desgaste más elevadas en las poblaciones muy móviles. Los encuestados a veces son compensados para minimizar la no respuesta, aunque la introducción de esa compensación tiene que ser estudiada detenidamente. En ocasiones, una vez que se han identificado todas las unidades que se deben seguir, se selecciona aleatoriamente una submuestra de estas unidades para un seguimiento muy intensivo, que puede requerir esfuerzos adicionales o alguna forma de compensación. En cualquier caso, el contrato para la agencia recopiladora de datos debe contener incentivos claros, como una mayor compensación si la tasa de no respuesta se mantiene por debajo de un umbral aceptable.

Concepto clave

Las evaluaciones de impacto con las mejores prácticas intentan mantener la no respuesta y el desgaste en el nivel más bajo posible.

Se deben establecer *procedimientos de garantía de calidad* bien definidos para todas las etapas de la recopilación de datos, incluyendo el diseño del procedimiento del muestreo y el cuestionario, las etapas de preparación, recopilación de datos, ingreso de los datos, y limpieza y almacenamiento de los mismos.

Se debería otorgar una gran prioridad a los controles de calidad durante el trabajo de campo, con el fin de minimizar los errores de cada unidad. Deben existir procedimientos claros para volver a visitar las unidades que no han proporcionado información o que han proporcionado información incompleta. Deben introducirse múltiples filtros en el proceso de control de calidad, por ejemplo, contando con encuestadores, supervisores y, si fuera necesario, coordinadores del trabajo de campo para que vuelvan a visitar a las unidades que no respondieron para verificar su estatus. Los cuestionarios de las entrevistas con no respuesta deberían ser codificados con claridad y registrados. Una vez que los datos han sido completamente digitalizados, las tasas de no respuesta se pueden resumir y se puede dar cuenta de todas las unidades de la muestra.

También deberían realizarse controles de calidad de cualquier dato incompleto para una unidad encuestada en particular. Una vez más, el proceso de control de calidad debe incluir múltiples filtros. El encuestador es el responsable de verificar los datos inmediatamente después de que han sido

recopilados. El supervisor y el coordinador del trabajo de campo deben llevar a cabo controles aleatorios en una etapa posterior.

Los controles de calidad de los errores de medición son más difíciles pero cruciales para evaluar si la información se ha recopilado con precisión. Los controles de consistencia se pueden incorporar en el cuestionario. Además, los supervisores o controladores de calidad tienen que llevar a cabo *controles in situ,* participando en las entrevistas para asegurar que los encuestadores recopilen los datos según los estándares de calidad establecidos. Se pueden realizar *verificaciones externas* o auditorías de calidad con una submuestra de la encuesta de evaluación de impacto para asegurar que los datos recopilados sean precisos. Esto a veces se lleva a cabo con un controlador de calidad que recoge un subconjunto del cuestionario con un encuestado y compara las respuestas con aquellas obtenidas anteriormente por un encuestador con el mismo encuestado.

Los coordinadores del trabajo de campo o los miembros del equipo de evaluación también deberían contribuir con los controles de calidad para minimizar los conflictos de interés potenciales en la empresa encuestadora. Puede que también sea necesario contratar una agencia externa para auditar la calidad de las actividades de recopilación de datos. Esto puede limitar significativamente la gama de problemas que puedan surgir debido a la falta de supervisión del equipo de recopilación de datos, o debido a procedimientos insuficientes de control de calidad.

En definitiva, es crucial que todos los pasos que intervienen en el control de calidad se realicen explícitamente en los términos de referencia cuando se contrata la recopilación de datos.

Procesamiento y almacenamiento de los datos

El procesamiento y la validación de los datos es una parte integral de la recopilación de datos de una nueva encuesta. Incluye los pasos para digitalizar la información de las encuestas de papel y lápiz, así como los pasos para validar los datos tanto de estas últimas como de la recopilación electrónica mediante ordenadores portátiles, teléfonos inteligentes, *Tablets* u otros instrumentos. Al trabajar con encuestas de papel y lápiz, se debe elaborar un *programa de entrada de datos* y se debe instaurar un sistema para gestionar el flujo de datos que serán digitados. Hay que establecer normas y procedimientos y capacitar rigurosamente a los operadores de ingreso de datos para garantizar que dicho ingreso sea consistente. En la medida de lo posible, el ingreso de datos debería ser integrado en las operaciones de recopilación de datos (incluida la fase de prueba piloto), de manera que cualquier problema con los datos recopilados se pueda identificar rápidamente y verificar en el terreno. En general, la referencia de calidad para el proceso de entrada de datos

debería ser que los datos físicos brutos fuesen replicados con exactitud en la versión digitalizada, sin modificaciones mientras se ingresan. Para minimizar los errores de ingreso de datos, se puede utilizar un procedimiento de *ingreso de datos de doble ciego*, de modo de identificar y corregir cualquier error adicional. Se puede aplicar un enfoque de entrada de campo asistida por ordenador, que recopila los datos en una encuesta de lápiz y papel y luego los digitaliza en el terreno y los valida de inmediato para identificar errores e inconsistencias.

Tanto en las encuestas de papel y lápiz como en las encuestas que dependen de la recopilación electrónica de datos, se pueden desarrollar programas para llevar a cabo controles automáticos de los errores no muestrales (tanto en las no respuestas como en inconsistencias de las entradas) que se pueden producir en el terreno, y para validar los datos. Si el proceso de validación se integra en los procedimientos del trabajo de campo, se pueden devolver los datos incompletos o inconsistentes a los trabajadores en el terreno para una verificación in situ. Este tipo de integración no está exenta de dificultades en lo que se refiere al flujo organizacional de las operaciones del trabajo de campo, pero puede producir importantes mejoras de la calidad, disminuir el error de medición y aumentar la potencia estadística de la evaluación de impacto. La posibilidad de utilizar un enfoque integrado de este tipo debe contemplarse explícitamente cuando se planifica la recopilación de datos. El uso de nuevas tecnologías puede facilitar esos controles de calidad.

Como ya se ha señalado, la recopilación de datos comprende un conjunto de operaciones cuya complejidad no debería ser subestimada. El recuadro 16.5 trata de cómo el proceso de recopilación de datos para la evaluación de las pruebas piloto de atención a crisis en Nicaragua produjo datos de alta calidad con muy bajo desgaste y pocas no respuestas a las preguntas, así como pocos errores de medición y de procesamiento. Estos datos de alta calidad se pueden obtener solo cuando se establecen los procedimientos de calidad de los datos y los incentivos adecuados al contratar la recopilación de datos.

Al final del proceso de recopilación, los datos deben presentarse con documentación detallada, lo que incluye un libro de códigos completo y un diccionario de datos, y deben almacenarse en un sitio seguro (véase el recuadro 16.6). Si los datos están siendo recopilados para una evaluación de impacto, el conjunto de datos también debe incluir información complementaria sobre la condición de tratamiento y la participación en el programa. Un paquete completo de documentación acelerará el análisis de los datos de evaluación de impacto, contribuirá a producir resultados que se pueden utilizar para la elaboración de las políticas de manera oportuna y facilitará la distribución de la información y la potencial replicación.

Recuadro 16.5: Recopilación de datos para la evaluación de las pruebas piloto de atención a crisis en Nicaragua

En 2005 el gobierno nicaragüense lanzó el programa piloto Atención a Crisis. Se elaboró un estudio para evaluar el impacto de combinar un programa de transferencias condicionadas con transferencias productivas, como ayudas para invertir en actividades no agrícolas o en formación profesional. La prueba piloto de Atención a Crisis fue implementada por el Ministerio de la Familia, con apoyo del Banco Mundial.

En la evaluación se utilizó una asignación aleatoria en dos etapas. En primer lugar, se asignaron 106 comunidades ya sea al grupo de comparación o al grupo de tratamiento. En segundo lugar, en las comunidades de tratamiento los hogares elegibles fueron asignados aleatoriamente a uno de tres paquetes de beneficios: transferencias condicionadas, la transferencia más una ayuda que permitía a uno de los miembros del hogar elegir entre diversos cursos de formación profesional, y transferencias monetarias más una ayuda para la inversión productiva, destinada a estimular a los receptores para el inicio de una actividad no agrícola, con el fin de crear activos y diversificar el ingreso (Macours, Premand y Vakis, 2012).

En 2005 se llevó a cabo una encuesta de línea de base, en 2006 se produjo una primera encuesta de seguimiento y en 2008 se realizó una segunda encuesta de seguimiento, dos años después de finalizada la intervención. Se establecieron rigurosos controles de calidad en todas las etapas del proceso de recopilación de datos. Primero, los cuestionarios fueron sometidos a una exhaustiva prueba en el terreno y se capacitó a los encuestadores tanto en las aulas como en las condiciones en el terreno. Segundo, se estableció una supervisión en el terreno de modo que todos los cuestionarios fueron revisados varias veces por los encuestadores, supervisores, coordinadores del trabajo de campo y otros expertos. Tercero, se utilizó un sistema de ingreso de datos de doble ciego, junto con un programa exhaustivo de control de calidad que podía identificar los cuestionarios incompletos o inconsistentes. Los cuestionarios donde faltaba información en ciertas preguntas o donde se observaban inconsistencias eran devueltos sistemáticamente al terreno para ser verificados. Estos procedimientos y requisitos fueron especificados de forma explícita en los términos de referencia de la empresa de recopilación de datos.

Además, se establecieron procedimientos detallados de seguimiento para minimizar el desgaste. Al comienzo, en 2008, se llevó a cabo un censo de los hogares que residían en las comunidades de tratamiento y de control, en estrecha colaboración con los dirigentes comunitarios. Dado que la migración en el país era habitual, a la empresa encargada de la encuesta se le ofrecieron incentivos para hacer un seguimiento de los migrantes individuales en todo el país. Como consecuencia, solo el 2% de los 4.359 hogares originales no pudieron ser entrevistados en 2009. La empresa de la encuesta también se encargó de dar seguimiento a todos los individuos de los hogares encuestados en 2005. Una vez más, solo no se pudo realizar el seguimiento de un 2% de los individuos objeto de las transferencias del programa (otro 2% había fallecido). El desgaste fue de un 6% para todos los niños de los hogares encuestados en 2005 y de un 5% para todos los individuos en los hogares encuestados en ese mismo año.

Continúa en la página siguiente.

Recuadro 16.5: Recopilación de datos para la evaluación de las pruebas piloto de atención a crisis en Nicaragua *(continúa)*

Las tasas de desgaste y de no respuesta proporcionan un buen indicador de la calidad de la encuesta. Las tasas de desgaste muy bajas requieren grandes esfuerzos de la empresa de recopilación de datos, así como incentivos explícitos. El costo unitario de un hogar o individuo objeto de un seguimiento también es mucho mayor. Además, en este caso, los controles de calidad rigurosos añadieron costos y aumentaron el tiempo de recopilación de datos. Aun así, en el contexto de la prueba piloto de Atención a Crisis, la muestra siguió siendo representativa tanto a nivel de los hogares como de los individuos tres a cuatro años después de la línea de base; se minimizaron los errores de medición, y se garantizó la fiabilidad de la evaluación. Como consecuencia, los impactos de largo plazo de las pruebas piloto de Atención a Crisis pudieron ser analizados de manera convincente.

Fuente: Macours, Premand y Vakis (2012).

Recuadro 16.6: Directrices para la documentación y el almacenamiento de datos

La práctica clave en la documentación de datos consiste en mantener un registro de todos los datos de la evaluación de impacto. Esto implica los protocolos de recopilación de datos, los cuestionarios, los manuales de formación y otros. El Banco Mundial, el Banco Interamericano de Desarrollo (BID) y la Millenium Challenge Corporation, entre otros organismos, tienen iniciativas de datos abiertos que ponen estos datos a disposición del público mediante un catálogo de datos.

El almacenamiento se puede descomponer en tres categorías: microdatos, macrodatos y archivos de control de identidad.

- Los *microdatos* son datos al nivel de la unidad de observación, que permanece anónima y no incluye ninguna información que identifique a los individuos. Las variables de identificación relevantes guardan el anonimato de la identificación, que está vinculada solo a la información de los encuestados en los ficheros de control de identidad.
- Los archivos *de control de identidad* contienen toda la información antes de que se vuelva anónima. Deben guardarse solo en un servidor seguro y nunca incluirse en un catálogo de datos.
- Los *macrodatos* comprenden todos los documentos de apoyo relevantes para la interpretación de los microdatos, el diccionario de datos, el libro de códigos, la descripción del diseño del estudio y los cuestionarios.

La catalogación de los macrodatos y microdatos contribuye a proteger la seguridad de los datos y cumple las normas internacionales sobre almacenamiento de datos. Los catálogos de los datos centrales son mucho menos vulnerables al mal funcionamiento o a la intrusión que el disco duro de un computador o un instrumento portátil de almacenamiento. En ciertos catálogos de datos, los datos pueden permanecer protegidos por una contraseña durante un período determinado antes de estar disponibles al público.

Otros recursos

- Para material de apoyo relacionado con el libro y para hipervínculos de más recursos, se recomienda consultar el sitio web de la Evaluación de Impacto en la Práctica (http://www.worldbank.org/ieinpractice).
- Para una guía del diseño del cuestionario, véase el módulo sobre "Técnicas del trabajo de campo aplicadas" en el curso de métodos de evaluación de impacto de la Universidad de California (http://aie.cega.org).
- Para entradas en los blogs sobre recopilación de datos, véase la lista documentada del blog de impacto en el desarrollo del Banco Mundial (http://blogs.worldbank.org/impactevaluations).
- Para más información sobre la recopilación de datos, véase el siguiente material:
  - ➢ A. G. Fink y J. Kosecoff (2008), *How to Conduct Surveys: A Step by Step Guide*, cuarta edición. Londres: Sage.
  - ➢ G. Iarossi (2006), *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, D.C.: Banco Mundial.
  - ➢ E. Leeuw, J. Hox y D. Dillman (2008), *International Handbook of Survey Methodology.* Nueva York: Taylor & Francis Group.
- Para más información sobre las actividades de recopilación de datos y supervisión de calidad de los datos, véase el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial, Módulo 5 sobre recopilación de datos (http://www.worldbank.org/health/impactevaluationtoolkit). El módulo incluye varios ejemplos de informes de progreso de la encuesta, manuales para el trabajo de campo y programas de capacitación para los hogares y centros de salud.
- Para diversos materiales de orientación y preparación de una encuesta, véase el portal de evaluación del Banco Interamericano de Desarrollo (http://www.iadb.org/portalevaluacion). En la sección de recopilación de datos, puede descargarse:
  - ➢ Un manual para el diseño del cuestionario.
  - ➢ Un manual de ingreso de datos.
  - ➢ Formularios de consentimiento, cuestionarios de la muestra, programas de ingreso de datos y manuales para el trabajo de campo en diferentes tipos de encuestas, entre ellas encuestas de hogares, de comunidades, de centros de salud, escuelas y agricultores.
  - ➢ Enlaces con otros ejemplos de preguntas y cuestionarios de las encuestas.
  - ➢ Enlaces con directrices para la recopilación de datos de calidad.
  - ➢ Enlaces con instrumentos disponibles en el sitio web de International Household Survey Network (IHSN) para almacenamiento y gestión de datos.
- Para más información sobre las razones de la importancia de la documentación de datos, cómo se puede llevar a cabo y quién es el responsable de ello en el equipo de evaluación, véase el Kit de Herramientas de Evaluación de Impacto (*Impact Evaluation Toolkit*) del Banco Mundial, Módulo 6, sobre almacenamiento de datos (http://www.worldbank.org/health/impactevaluationtoolkit).

Notas

1. Véanse referencias en Grosh y Glewwe (2000) y Naciones Unidas (2005). Véanse también Muñoz (2005); Iarossi (2006); Fink y Kosecoff (2008), y Leeuw, Hox y Dillman (2008), que proporcionan abundante orientación práctica sobre la recopilación de datos.
2. Consúltese McKenzie y Rosenzweig (2012) para una visión general de los últimos avances.
3. Para ejemplos de este tipo de experimentos, véanse McKenzie y Rosenzweig (2012) en temas generales; Beegle, Carletto y Himelein (2012) sobre datos agrícolas; Beegle et al. (2012) sobre la medición del consumo de los hogares, y Bardasi et al. (2011) sobre datos laborales.
4. Para ejemplos de innovaciones en la medición de resultados, véase Holla (2013); Das y Hammer (2007), y Planas et al. (2015).

Referencias bibliográficas

Baird, S. y B. Özler. 2012. "Examining the Reliability of Self-reported Data on School Participation." *Journal of Development Economics* 98 (1): 89–93.

Bardasi, E., K. Beegle, A. Dillon, A. y P. Serneels. 2011. "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania." *The World Bank Economic Review* 25 (3): 418–47.

Beegle, K., C. Carletto y K. Himelein. 2012. "Reliability of Recall in Agricultural Data." *Journal of Development Economics* 98 (1): 34–41.

Beegle, K., J. De Weerdt, J. Friedman y J. Gibson. 2012. "Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania." *Journal of Development Economics* 98 (1): 3–18.

Caeyers, B., N. Chalmers y J. De Weerdt. 2012. "Improving Consumption Measurement and Other Survey Data through CAPI: Evidence from a Randomized Experiment." *Journal of Development Economics* 98 (1): 19–33.

Chetty, R., J. N. Friedman y E. Sáez. 2013. "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review* 103 (7): 2683–2721.

Das, J. y J. Hammer. 2007. "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India." *Journal of Development Economics* 83 (1): 1–36.

Fafchamps, M., D. McKenzie, S. Quinn y C. Woodruff. 2012. "Using PDA Consistency Checks to Increase the Precision of Profits and Sales Measurement in Panels." *Journal of Development Economics* 98 (1): 51–57.

Ferraz, C. y F. Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *The Quarterly Journal of Economics* 123 (2): 703–45.

Fink, A. G. y J. Kosecoff. 2008. *How to Conduct Surveys: A Step by Step Guide*, cuarta edición. Londres: Sage.

Galiani, S., P. Gertler y E. Schargrodsky, E. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Galiani, S. y P. McEwan. 2013. "The Heterogeneous Impact of Conditional Cash Transfers." *Journal of Public Economics* 103: 85–96.

Gertler, P., P. Giovagnoli y S. Martínez. 2014. "Rewarding Provider Performance to Enable a Healthy Start to Life: Evidence from Argentina's Plan Nacer." Documento de trabajo de investigación de políticas Núm. 6884. Washington, D.C.: Banco Mundial.

Glewwe, P. 2005. "An Overview of Questionnaire Design for Household Surveys in Developing Countries." En: *Household Sample Surveys in Developing and Transition Countries*. Nueva York: Naciones Unidas.

Glewwe, P. y P. Olinto. 2004. "Evaluating the Impact of Conditional Cash Transfers on Schooling: An Experimental Analysis of Honduras' PRAF Program." Informe final. University of Minnesota y IFPRI-FCND.

Grosh, M. y P. Glewwe (eds.). 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study.* Washington, D.C.: Banco Mundial.

Holla, A. 2013. "Measuring the Quality of Health Care in Clinics." Washington, D.C.: Banco Mundial. Disponible en http://www.globalhealthlearning.org/sites/default/files/page-files/Measuring%20Quality%20of%20Health%20Care_020313.pdf.

Iarossi, G. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, D.C.: Banco Mundial.

Kasprzyk, D. 2005. "Measurement Error in Household Surveys: Sources and Measurement." En: *Household Sample Surveys in Developing and Transition Countries*. Nueva York: Naciones Unidas.

Leeuw, E., J. Hox y D. Dillman. 2008. *International Handbook of Survey Methodology.* Nueva York: Taylor & Francis Group.

Macours, K., P. Premand y R. Vakis. 2012. "Transfers, Diversification and Household Risk Strategies: Experimental Evidence with Implications for Climate Change Adaptation." Documento de trabajo de investigación de políticas Núm. 6053. Washington, D.C.: Banco Mundial.

McKenzie, D. y M. Rosenzweig. 2012. "Symposium on Measurement and Survey Design." *Journal of Development Economics* 98 (1, Mayo): 1–148.

Morris, S. S., R. Flores, P. Olinto y J. M. Medina. 2004. "Monetary Incentives in Primary Health Care and Effects on Use and Coverage of Preventive Health Care Interventions in Rural Honduras: Cluster Randomized Trial." *Lancet* 364: 2030–37.

Muñoz, J. 2005. "A Guide for Data Management of Household Surveys." En: *Household Sample Surveys in Developing and Transition Countries*. Nueva York: Naciones Unidas.

Naciones Unidas. 2005. *Household Sample Surveys in Developing and Transition Countries.* Nueva York: Naciones Unidas.

Planas, M-E, P. J. García, M. Bustelo, C. P. Cárcamo, S. Martínez, H. Ñopo, J. Rodríguez, M. F. Merino y A. Morrison. 2015. "Effects of Ethnic Attributes on the Quality of Family Planning Services in Lima, Peru: A Randomized Crossover Trial." *PLoS ONE* 10 (2): e0115274.

Pradhan, M. y L. B. Rawlings. 2002. "The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund." *World Bank Economic Review* 16 (2): 275–95.

Vermeersch, C., E. Rothenbühler y J. Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. Washington, D.C.: Banco Mundial. Disponible en http://www.worldbank.org/health/impactevaluationtoolkit.

CAPÍTULO 17

Conclusiones

Las evaluaciones de impacto: ejercicios complejos pero valiosos

La evaluación de impacto tiene que ver con generar evidencia sobre qué programas funcionan, qué programas no lo hacen y cómo mejorarlos para lograr mejores resultados en materia de desarrollo. Esto se puede realizar en un marco clásico de evaluación de impacto, contrastando los resultados entre grupos de tratamiento y comparación. Las evaluaciones de impacto también se pueden llevar a cabo para explorar alternativas de implementación de un programa, para probar innovaciones o analizar diferentes programas con el fin de evaluar el desempeño comparativamente.

La idea que subyace a este trabajo es que las evaluaciones de impacto constituyen una inversión valiosa para numerosos programas. Junto con el monitoreo y otras formas de evaluación, optimizan la comprensión de la efectividad de determinadas políticas; contribuyen a una rendición de cuentas mejorada de los administradores del programa, los gobiernos, los financiadores y el público en general; fundamentan decisiones acerca de cómo asignar de manera más eficiente los escasos recursos para el desarrollo, y aportan al acervo global de conocimientos sobre lo que funciona y no funciona en el campo del desarrollo.

Lista de verificación: elementos centrales de una evaluación de impacto bien diseñada

Las evaluaciones de impacto son ejercicios complejos con numerosas partes cambiantes. La siguiente lista de verificación destaca los elementos centrales de una evaluación de impacto bien diseñada:

- ✓ Una pregunta concreta y relevante para las políticas –basada en una teoría del cambio– a la que se puede responder con una evaluación de impacto.
- ✓ Una metodología robusta, derivada de las reglas operativas del programa, para estimar un contrafactual que muestre la relación causal entre el programa y los resultados de interés.
- ✓ Un equipo de evaluación bien formado que funcione como la asociación de un equipo de políticas públicas y un equipo de investigación.
- ✓ Respeto por las normas éticas y consideración por los sujetos humanos en el diseño y la implementación de la evaluación y la recopilación de datos correspondiente, así como atención a los principios de ciencia abierta para asegurar la transparencia.
- ✓ Una muestra con suficiente potencia estadística para permitir que se detecten los impactos relevantes para las políticas.
- ✓ Una metodología y una muestra que proporcionen resultados generalizables para la población de interés.
- ✓ Datos de gran calidad que proporcionen la información adecuada requerida para la evaluación de impacto, incluidos los datos de los grupos de tratamiento y de comparación, los datos de línea de base y de seguimiento, y la información sobre la implementación y los costos del programa.
- ✓ Una estrategia de participación para fundamentar el diálogo de políticas a través de la implementación de la evaluación de impacto, así como también un informe de evaluación de impacto y notas informativas de las políticas relacionadas divulgadas al público objetivo en el momento oportuno.

Lista de verificación: recomendaciones para mitigar riesgos habituales al llevar adelante una evaluación de impacto

También se destacan algunas recomendaciones que pueden contribuir a mitigar los riesgos habituales inherentes al proceso de realización de una evaluación de impacto:

- ✓ El mejor momento para diseñar una evaluación de impacto es temprano en el ciclo del proyecto, idealmente como parte del diseño del programa, pero al menos antes de implementar el programa que se evalúa. Una planificación temprana permite un diseño de evaluación prospectivo basado en la mejor metodología disponible, y brindará el tiempo necesario para planificar e implementar la recopilación de datos de línea de base en los ámbitos de la evaluación antes de que comience el programa.

- ✓ Los resultados de la evaluación de impacto deben acompañarse con evaluaciones complementarias del proceso y datos de monitoreo que muestren un cuadro claro de la implementación del programa. Cuando los programas tienen éxito, es importante entender por qué. Cuando los programas fracasan, es importante poder distinguir entre un programa mal implementado y un diseño de programa deficiente.

- ✓ Se deben recopilar los datos de línea de base e incorporar una metodología de respaldo en el diseño de la evaluación de impacto. Si el diseño de la evaluación original es invalidado –por ejemplo, porque el grupo de comparación original recibe los beneficios del programa–, contar con un plan de respaldo puede ayudar a evitar tener que renunciar por completo a la evaluación.

- ✓ Se deben mantener identificadores comunes entre diferentes fuentes de datos para las unidades de observación, de modo que se les pueda relacionar fácilmente durante el análisis. Por ejemplo, un determinado hogar debe tener el mismo identificador en los sistemas de monitoreo y en las encuestas de línea de base y de seguimiento de la evaluación de impacto.

- ✓ Las evaluaciones de impacto son útiles para aprender cómo funcionan los programas y para probar alternativas de programas, incluso en el caso de grandes programas en curso. Las evaluaciones de impacto bien

diseñadas pueden contribuir a probar innovaciones o a proporcionar nuevas perspectivas de la efectividad relativa de diversos productos y servicios prestados como un paquete de programas existentes. Incorporar una innovación adicional al programa como un pequeño plan piloto en el contexto de una evaluación más grande puede servir para aprovechar la evaluación a fin de producir información valiosa para la toma de decisiones en el futuro.

- ✓ Se debe pensar en las evaluaciones de impacto como otro de los componentes de la operación de un programa y se les debe dotar de personal adecuado y de un presupuesto que contemple los recursos técnicos y financieros necesarios. Es preciso ser realista acerca de los costos y la complejidad de llevar a cabo una evaluación de impacto. Normalmente, el proceso de diseñar una evaluación y recopilar una línea de base desde cero puede tardar un año o más. Una vez que el programa comienza, el grupo de tratamiento necesita un período suficientemente largo de exposición a la intervención para influir en los resultados. Dependiendo del programa, esto puede tardar entre un año y cinco años, o más en el caso de resultados de largo plazo. Recopilar una o más encuestas de seguimiento, llevar a cabo el análisis y divulgar los resultados también requerirá un esfuerzo considerable a lo largo de varios meses y años. En su conjunto, un ciclo completo de evaluación de impacto desde el comienzo hasta el final suele durar entre tres y cuatro años de intensivo trabajo y participación. Se requieren recursos financieros y técnicos adecuados en cada paso del proceso.

Eventualmente, las evaluaciones de impacto individuales ofrecen respuestas concretas a preguntas específicas de políticas públicas. Aunque estas respuestas proporcionan información hecha a la medida de la entidad específica que encarga y financia la evaluación, también brindan información valiosa para otros agentes en otras partes del mundo, que pueden aprender y tomar decisiones sobre la base de la evidencia. Por ejemplo, los programas de transferencias condicionadas de África, Asia y Europa han extraído enseñanzas de las evaluaciones originales de Familias en Acción de Colombia, Progresa de México y otros programas de transferencias condicionadas de América Latina. De esta manera, las evaluaciones de impacto constituyen en parte un bien público global. La evidencia que se genera a través de una evaluación de impacto se suma al conocimiento mundial sobre este tema. Esta base de conocimientos luego puede fundamentar decisiones de políticas de otros países y contextos, prestando la atención adecuada a la validez externa. La comunidad internacional ha avanzado rápidamente hacia un apoyo de mayor escala de evaluaciones rigurosas.

A nivel de país, gobiernos cada vez más sofisticados y exigentes esperan demostrar resultados y ser más capaces de rendir cuentas ante sus electores clave. Se emprenden cada vez más evaluaciones de impacto de la mano de los ministerios nacionales y subnacionales pertinentes, y los órganos de gobierno creados para dirigir una agenda nacional de evaluación, como el Consejo Nacional de Evaluación de la Política de Desarrollo Social, en México, y el Departamento de Monitoreo y Evaluación del Desempeño en Sudáfrica (*Department of Performance Monitoring and Evaluation*). También se utiliza la evidencia de estas evaluaciones para fundamentar las asignaciones presupuestarias propuestas por el Congreso y el Parlamento a nivel nacional. En los sistemas donde los programas se juzgan a partir de la evidencia y los resultados finales, los programas que tienen una sólida base de evidencia para defender resultados positivos podrán salir adelante, mientras que los que carecen de dichas pruebas tendrán más dificultades para encontrar financiamiento.

Las instituciones multilaterales como el Banco Mundial y el Banco Interamericano de Desarrollo (BID), así como los organismos nacionales de desarrollo, los gobiernos donantes y las instituciones filantrópicas también exigen más y mejor evidencia sobre el uso efectivo de los recursos para el desarrollo. Esta evidencia se requiere para rendir cuentas a quienes prestan o donan el dinero, y para la toma de decisiones acerca de dónde es mejor asignar los escasos recursos para el desarrollo.

Asimismo, está emergiendo un número creciente de instituciones dedicadas principalmente a la producción de evaluaciones de impacto de alta calidad, entre ellas las del ámbito académico como Poverty Action Lab (J-Pal), Innovations for Poverty Action (IPA), y el Center for Effective Global Action (CEGA), y organismos independientes que apoyan las evaluaciones de impacto, como la Iniciativa Internacional para la Evaluación de Impacto (3ie). Hay diversas asociaciones que reúnen a grupos de profesionales de la evaluación e investigadores y responsables de las políticas interesados en el tema, entre ellas la Network of Networks on Impact Evaluation y asociaciones regionales como la African Evaluation Association y la Red de Evaluación de Impacto de la Asociación Económica de América Latina y el Caribe. Todas estas iniciativas reflejan la creciente importancia de la evaluación de impacto en las políticas internacionales de desarrollo.

Debido a estos avances, poder comunicarse en el lenguaje de la evaluación de impacto es una habilidad cada vez más indispensable para cualquier profesional del desarrollo, ya sea para quienes se ganan la vida trabajando en evaluaciones, o bien para los que contratan evaluaciones de impacto o utilizan los resultados de las mismas en la toma de decisiones. La rigurosa evidencia generada a través de las evaluaciones de impacto puede ser uno de los

motores del diálogo de políticas para el desarrollo, y proporcionar la base para apoyar o para oponerse a las inversiones en programas y políticas de desarrollo. La evidencia de las evaluaciones de impacto permite a los responsables de las políticas y a los administradores de los proyectos tomar decisiones fundamentadas sobre cómo alcanzar resultados de la manera más costo-efectiva. Armado con la evidencia de una evaluación de impacto, el equipo de políticas públicas tiene el deber de cerrar el ciclo incorporando esos resultados en el proceso de toma de decisiones. Este tipo de evidencia puede respaldar debates, opiniones y, en definitiva, las decisiones de asignación de recursos humanos y monetarios de los gobiernos, las instituciones multilaterales y los donantes.

La elaboración de políticas basadas en la evidencia tiene que ver esencialmente con fundamentar el diseño de programas y mejorar la asignación presupuestaria para ampliar los programas costo-efectivos, eliminar los inefectivos e introducir mejoras en los diseños sobre la base de la mejor evidencia disponible. Las evaluaciones de impacto no son una empresa puramente académica. Son el resultado de la necesidad de encontrar respuestas a las preguntas de políticas que influyen en la vida diaria de las personas. Las decisiones sobre cuál es la mejor manera de asignar recursos escasos a los programas de lucha contra la pobreza, o de transporte, energía, salud, educación, de redes de protección, microcréditos, agricultura, y otras innumerables iniciativas para el desarrollo, tienen el potencial para mejorar el bienestar de las personas en todo el mundo. Es vital que esas decisiones se tomen utilizando la evidencia más rigurosa posible.

GLOSARIO

Los términos que llevan cursiva dentro de las definiciones se definen a su vez dentro del propio glosario.

Actividad. Medidas adoptadas o trabajo realizado a través del cual los *insumos*, como los fondos, la asistencia técnica y otro tipo de recursos que se movilizan para generar *productos* específicos, como el dinero gastado, los libros de texto distribuidos o el número de participantes en un programa de empleo.

Análisis de costo-beneficio. Estima los beneficios totales previstos de un programa, en comparación con sus costos totales previstos. Su fin es cuantificar todos los costos y beneficios de un programa en términos monetarios y evaluar si los beneficios superan a los costos.

Análisis de costo-efectividad. Compara el costo relativo de dos o más programas o alternativas de programa en términos de alcanzar un resultado común, como la producción agrícola o las calificaciones de los alumnos en los exámenes.

Análisis de regresión. Método estadístico para analizar las relaciones entre una *variable dependiente* (la variable que se debe explicar) y *variables explicativas*. El análisis de regresión normalmente no es suficiente para capturar los efectos causales. En la *evaluación de impacto*, el análisis de regresión es una manera de representar la relación entre el valor de un indicador de *resultado Y* (variable dependiente) y una variable independiente que captura la asignación al *grupo de tratamiento* o *grupo de comparación*, mientras se mantienen constantes otras características. Tanto la asignación al grupo de tratamiento y de comparación como las demás características son variables explicativas. El análisis de regresión puede ser univariante (si hay solo una variable explicativa; en el caso de la evaluación de impacto, la única variable explicativa es la asignación al grupo de tratamiento o de comparación) o multivariante (si hay varias variables explicativas).

Análisis de sensibilidad. Trata de la sensibilidad del análisis ante los cambios en los supuestos. En el contexto de los *cálculos de potencia*, contribuye a comprender

cuánto tendrá que aumentar el tamaño requerido de la *muestra* bajo supuestos más conservadores (como un menor impacto esperado, una mayor variación en el indicador de resultado o un nivel más alto de *potencia*).

Asignación aleatoria o ensayo controlado aleatorio. Método de *evaluación de impacto* por el cual cada unidad elegible (por ejemplo, un individuo, un hogar, una empresa, una escuela, un hospital o una comunidad) tiene la misma probabilidad de ser seleccionada para ser tratada en un programa. Con un número suficientemente grande de unidades, el proceso de asignación aleatoria garantiza la equivalencia tanto en las características observables como no observables entre el *grupo de tratamiento* y el *grupo de comparación*, y así se descarta cualquier *sesgo de selección*. La asignación aleatoria se considera el método más robusto para estimar los *contrafactuales* y se le suele considerar como la regla de oro de la *evaluación de impacto*.

Cadena de resultados. Establece la lógica causal del programa explicando cómo se logra el objetivo de desarrollo. Articula la secuencia de *insumos, actividades* y *productos* que se espera que mejoren los resultados.

Cálculos de potencia. Método para determinar cuál es el tamaño de la *muestra* requerida para que una *evaluación de impacto* estime con precisión el impacto de un programa, es decir: la muestra más pequeña que permitirá detectar el *efecto mínimo detectable*. Los cálculos de potencia dependen de parámetros como la *potencia* (o la probabilidad de un *error de tipo II*), el *nivel de significancia*, la media, la varianza y la *correlación intra-clusters* de los *resultados* de interés.

Censo. Empadronamiento total de una población. Los datos censales abarcan todas las unidades de la población. Compárese con *muestra*.

Ciencia abierta. Movimiento cuyo fin es elaborar métodos de investigación más transparentes, mediante el registro de los ensayos, la utilización de planes de preanálisis, documentación de datos y registros.

Comparación antes y después. También conocida como "comparación previa-posterior" o "comparación reflexiva". Se trata de un seguimiento de los cambios en los *resultados* para los beneficiarios del programa a lo largo del tiempo, utilizando mediciones antes y después de la implementación del programa o la política, sin utilizar un *grupo de comparación*.

Comparaciones de inscritos y no inscritos. También conocidas como comparaciones autoseleccionadas. Esta estrategia compara los resultados de las *unidades* que decidieron inscribirse y las *unidades* que decidieron no inscribirse en un programa.

Cluster. También llamado conglomerado. Grupo de unidades que pueden compartir características similares. Por ejemplo, los niños que asisten a la misma escuela pertenecen a un mismo *cluster* porque comparten el mismo centro escolar, los mismos profesores y viven en el mismo barrio.

Consentimiento informado. Uno de los fundamentos de la protección de los derechos de los sujetos humanos. En el caso de las *evaluaciones de impacto*, requiere que los encuestados comprendan claramente los fines, procedimientos, riesgos y beneficios de la recopilación de datos en la que se les pide participar.

Contrafactual. Valor que habría tenido el *resultado* (*Y*) para los participantes del programa si no hubieran participado en el programa (*P*). Por definición, el contrafactual no se puede observar. Por lo tanto, debe estimarse utilizando un *grupo de comparación*.

Correlación. Medida estadística que indica hasta qué punto dos o más *variables* fluctúan juntas.

Correlación intra-*clusters*. También conocida como "correlación intraclase". Se trata del nivel de similitud en los *resultados* o características entre las unidades de los grupos previamente existentes o *clusters* en relación con unidades de otros *clusters*. Por ejemplo, los niños que asisten a la misma escuela normalmente serían más similares o estarían más correlacionados en términos de sus zonas de residencia o antecedentes socioeconómicos, en comparación con niños que no asisten a esa escuela.

Cumplimiento. Fenómeno que se produce cuando las *unidades* adhieren a su asignación como parte del *grupo de tratamiento* o del *grupo de comparación*.

Cumplimiento imperfecto. Discrepancia entre el estatus de tratamiento asignado y la condición de tratamiento real. Se produce cuando algunas unidades asignadas al *grupo de comparación* participan en el programa, o cuando algunas unidades asignadas al *grupo de tratamiento* no participan.

Datos administrativos. Datos recopilados asiduamente por organismos públicos o privados como parte de la administración de un programa, normalmente con frecuencia periódica y a menudo en el lugar de la prestación de servicios, e incluyen los servicios prestados, los costos y la participación en el programa. Los *datos de monitoreo* constituyen un tipo de *datos administrativos*.

Datos de encuesta. Datos que cubren una *muestra* de la población de interés. Compárese con *censo*.

Datos de monitoreo. Datos provenientes del *monitoreo* del programa que proporcionan información esencial a propósito de la prestación de una *intervención*, e incluyen quiénes son los beneficiarios y qué beneficios o *productos* del programa pueden haber recibido. Los datos de monitoreo constituyen un tipo de *datos administrativos*.

Desgaste. El desgaste se produce cuando algunas unidades abandonan la *muestra* entre una ronda de datos y la siguiente. Por ejemplo, cuando las personas cambian su residencia y no se les puede localizar. El desgaste de la muestra es un caso de *falta de respuesta* de la unidad. Puede introducir un *sesgo* en la estimación de impacto.

Diferencias en diferencias. También conocido como "doble diferencia" o "DD". Las diferencias en diferencias comparan los cambios en los *resultados* a lo largo del tiempo entre el *grupo de tratamiento* y el *grupo de comparación*. Esto elimina cualquier diferencia entre estos grupos que sea constante a lo largo del tiempo.

Diseño cruzado. También denominado diseño transversal. Se produce cuando hay una *asignación aleatoria* con dos o más intervenciones, lo que permite estimar el impacto de las intervenciones individuales y combinadas.

Diseño de regresión discontinua (DRD). Método de *evaluación de impacto cuasi experimental* que se puede utilizar en programas que dependen de un índice

continuo para clasificar a los participantes potenciales. Dicho índice tiene un punto límite que determina si los participantes potenciales son elegibles para recibir el programa o no. El umbral de elegibilidad del programa proporciona un punto divisorio entre el *grupo de tratamiento* y el *grupo de comparación*. Los resultados para los participantes en un lado del umbral se comparan con los resultados de los no participantes al otro lado del umbral. Cuando todas las unidades cumplen con la asignación que corresponde sobre la base de su índice de elegibilidad, se dice que el DRD es "nítido". Si hay incumplimiento en el otro lado del umbral, se dice que el DRD es "difuso" o "borroso".

Efecto causal. Véase *impacto*.

Efecto de derrame. También denominado efecto de contagio. Ocurre cuando el grupo de tratamiento influye directa o indirectamente en los *resultados* del grupo de comparación (o a la inversa).

Efecto mínimo detectable. El efecto mínimo detectable es un insumo en los *cálculos de potencia*, es decir, proporciona el tamaño del efecto que una *evaluación de impacto* está diseñada para estimar con un determinado nivel de *significancia* y *potencia*. Las *muestras* de la evaluación tienen que ser lo bastante grandes para distinguir al menos el efecto mínimo detectable. Este efecto se determina teniendo en cuenta el cambio en los *resultados* que justificaría la inversión que se ha hecho en una *intervención*.

Efecto Hawthorne. Se produce cuando, por el simple hecho de ser observadas, las unidades se comportan de manera diferente.

Efecto John Henry. Acontece cuando las unidades de la comparación se esfuerzan más para compensar que no se les haya ofrecido el tratamiento. Cuando se comparan las unidades tratadas con las unidades de la comparación que se esfuerzan más, la estimación del impacto del programa tiene un *sesgo*; es decir, se estima un impacto menor del programa en comparación con el impacto real que se obtendría si las unidades de la comparación no realizaran un esfuerzo adicional.

Efecto local promedio del tratamiento (LATE, por sus siglas en inglés). *Impacto* de un programa estimado para un subconjunto específico de la población, como las *unidades* que cumplen con su asignación al grupo de tratamiento o de comparación en presencia de un *cumplimiento imperfecto*, o en torno al umbral de elegibilidad cuando se aplica un *diseño de regresión discontinua*. Por lo tanto, el LATE proporciona solo una estimación local del *impacto* del programa y no debería generalizarse al conjunto de la población.

Efectos de equilibrio de contexto. *Efectos de derrame* que se producen cuando una *intervención* influye en las normas de conducta o sociales en un contexto determinado, como una localidad tratada.

Efectos de equilibrio general. Estos *efectos de derrame* se producen cuando las *intervenciones* afectan la oferta y demanda de bienes y servicios y, por lo tanto, cambian el precio de mercado de esos servicios.

Efecto promedio del tratamiento (ATE, por sus siglas en inglés). Impacto del programa bajo el supuesto de *cumplimiento total*; es decir, todas las *unidades* que hayan sido asignadas a un programa se inscriben realmente en él, y ninguna de las unidades de comparación recibe el programa.

Encuesta de seguimiento. También conocida como encuesta "posterior a la intervención". Se trata de una encuesta realizada después de que el programa ha comenzado, una vez que los beneficiarios se han beneficiado de él durante algún tiempo. Una *evaluación de impacto* puede incluir varias encuestas de seguimiento, a veces denominadas encuestas "intermedias" y "finales".

Equipo de evaluación. Equipo que lleva a cabo la *evaluación*. Se trata en esencia de una asociación entre dos grupos: un equipo de responsables de las políticas públicas (equipo de políticas) y un equipo de investigadores (equipo de investigación).

EMARF (en inglés, SMART). Específico, medible, atribuible, realista y focalizado. Los buenos *indicadores* tienen estas características.

Error de tipo I. También conocido como falso positivo. Este error se comete cuando se rechaza una *hipótesis nula* que, en realidad, es válida. En el contexto de una *evaluación de impacto*, se comete un error de tipo I cuando una *evaluación* llega a la conclusión de que un programa ha tenido un *impacto* (es decir, se rechaza la hipótesis nula de falta de impacto), aunque en realidad el programa no ha tenido impacto (es decir, la hipótesis nula se mantiene). El *nivel de significancia* es la probabilidad de cometer un error de tipo I.

Error de tipo II. También conocido como falso negativo. Este error se comete cuando se acepta (no se rechaza) la *hipótesis nula*, pese a que esta última, de hecho, no es válida. En el contexto de una *evaluación de impacto*, se comete un error de tipo II cuando se llega a la conclusión de que un programa no ha tenido *impacto* (es decir no se rechaza la hipótesis nula de falta de impacto), aunque el programa de hecho tuvo impacto (es decir, la hipótesis nula no es válida). La probabilidad de cometer un error de tipo II es 1 menos el nivel de *potencia*.

Estimación de tratamiento en los tratados. La estimación del impacto del tratamiento en aquellas *unidades* que en la práctica se han beneficiado del tratamiento. Compárese con *intención de tratar*.

Estimador. En Estadística, es una regla utilizada para calcular una característica desconocida de una población a partir de los datos (técnicamente conocido como "parámetro"); una estimación es el resultado de la aplicación real de una regla a una muestra concreta de datos.

Estudio de efectividad. Analiza si un programa funciona en condiciones normales al aumentar la escala. Cuando están adecuadamente diseñados e implementados, los resultados de estos estudios pueden ser más generalizables que en el caso de los *estudios de eficacia*.

Estudio de eficacia. Analiza si un programa puede funcionar en condiciones ideales. Estos estudios se llevan a cabo en circunstancias muy específicas, a menudo con una fuerte participación técnica de los investigadores durante la implementación

del programa. Suelen emprenderse para probar la viabilidad de un programa nuevo. Sus resultados no son generalizables más allá del alcance de la evaluación.

Evaluación. Valoración periódica y objetiva de un proyecto, un programa o una política planificados, en curso o finalizados. Las evaluaciones se utilizan para responder preguntas específicas, a menudo relacionadas con el diseño, la implementación o los resultados.

Evaluación de impacto. *Evaluación* que intenta establecer un vínculo causal entre un programa o *intervención* y un conjunto de *resultados*. Una evaluación de impacto procura responder a la pregunta: ¿cuál es el *impacto* (o efecto causal) de un programa en un *resultado* de interés?

Evaluación de proceso. *Evaluación* que se centra en cómo se implementa y funciona un programa, y que evalúa si se corresponde con su diseño original y documenta su desarrollo y funcionamiento. Compárese con *evaluación de impacto*.

Evaluación prospectiva. Evaluaciones diseñadas y aplicadas antes de que se implemente un programa. Las evaluaciones prospectivas están incorporadas en los planes de implementación del programa. Compárese con *evaluación retrospectiva*.

Evaluación retrospectiva. Evaluación diseñada después de que se ha implementado un programa (ex post). Compárese con *evaluación prospectiva*.

Experimento de mecanismo. *Evaluación de impacto* que prueba un mecanismo causal particular dentro de la *teoría del cambio* de un programa, en lugar de probar el efecto causal (*impacto*) del programa como un todo.

Factor invariante en el tiempo. Factor que no varía a lo largo del tiempo; es constante.

Factor variante en el tiempo. Factor que varía a lo largo del tiempo.

Falta de rango común. Cuando se utiliza el método de *pareamiento*, la falta de rango común es una falta de superposición entre los *puntajes de propensión* del grupo de tratamiento, o inscrito, y los del grupo de no inscritos.

Falta de respuesta. Se produce cuando faltan datos o los datos son incompletos para algunas unidades de la muestra. La *falta de respuesta de la unidad* surge cuando no hay información disponible para algunas unidades de la *muestra*, es decir, cuando la *muestra* real es diferente de la *muestra* planificada. Una forma de falta de respuesta a nivel de la unidad es el *desgaste*. La *falta de respuesta de una entrada* se produce cuando los datos son incompletos para algunas unidades de la muestra en un determinado momento del tiempo. La falta de respuesta puede generar *sesgos* en los resultados de una *evaluación* si está asociada con la condición de tratamiento.

Falta de respuesta de la unidad. Surge cuando no hay información disponible para un subconjunto de unidades; es decir, cuando la muestra real es diferente de la muestra planificada.

Falta de respuesta de una entrada. Ocurre cuando los datos son incompletos para algunas unidades de la *muestra*.

Generabilidad. La medida en que los resultados de una *evaluación* en un ámbito local serán válidos en otros contextos y en otros grupos de población.

Grupo de comparación. También conocido como *grupo de control*. Un grupo de comparación válido tendrá las mismas características, en promedio, que el grupo de beneficiarios del programa (*grupo de tratamiento*), con la única diferencia de que las unidades del grupo de comparación no se benefician del programa que se evalúa. Los grupos de comparación se utilizan para estimar el *contrafactual*.

Grupo de control. También conocido como *grupo de comparación* (véase la definición).

Grupo de tratamiento. También conocido como grupo tratado o grupo de intervención. El grupo de tratamiento es el grupo de *unidades* que es objeto de una *intervención* versus el *grupo de comparación*, que no es objeto de ella.

Hipótesis. Explicación propuesta de un fenómeno observable. Véase también *hipótesis nula* e *hipótesis alternativa*.

Hipótesis alternativa. Suposición de que la *hipótesis nula* es falsa. En una *evaluación de impacto*, la hipótesis alternativa suele ser la *hipótesis* de que la intervención tiene un impacto en los *resultados*.

Hipótesis nula. *Hipótesis* que puede ser falsificada sobre la base de los datos observados. Normalmente, la hipótesis nula propone una posición general o por defecto. En la *evaluación de impacto*, la hipótesis nula suele ser que el programa no tiene *impacto,* es decir: la diferencia entre el resultado del *grupo de tratamiento* y el *grupo de comparación* es cero.

Impacto. También conocido como *efecto causal*. En el contexto de las *evaluaciones de impacto,* un impacto es un cambio directamente atribuible a un programa, a una modalidad del programa o a innovaciones de diseño.

Indicador. *Variable* que mide un fenómeno de interés para el evaluador. El fenómeno puede ser un *insumo*, un *producto,* un *resultado*, una característica o un atributo. Véase también *EMARF*.

Índice de elegibilidad. También conocido como variable forzada. Se trata de una *variable* que permite clasificar a la *población de interés* a lo largo de una línea continua y tiene un umbral o una puntuación límite que determina quién es elegible y quién no lo es.

Insumos. Los recursos financieros, humanos y materiales utilizados en la *intervención*.

Intención de tratar (ITT, por sus siglas en inglés). Las estimaciones de ITT miden la diferencia en los *resultados* entre las *unidades* asignadas al *grupo de tratamiento* y las asignadas al *grupo de comparación*, independientemente de si las unidades de cada grupo recibieron en realidad el tratamiento.

Intervención. En el contexto de la evaluación de impacto, se trata del proyecto, del programa o de la política que se evalúa. También conocida como *tratamiento*.

Junta de revisión institucional (JRI). Comité nombrado para examinar, aprobar y monitorear la investigación con sujetos humanos. También conocido como Comité de ética independiente o Junta de revisión ética.

Línea de base. Situación previa a una *intervención*, con respecto a la cual se puede valorar el progreso o se pueden hacer comparaciones. La línea de base se recopila en forma previa a la implementación de un programa o política para observar la situación *antes*. La disponibilidad de datos de línea de base es fundamental para documentar el equilibrio en las características anteriores al programa entre los grupos de tratamiento y de comparación. Los datos de línea de base son necesarios para algunos diseños *cuasi experimentales*.

Marco muestral. Lista exhaustiva de las unidades de la *población de interés*. Se requiere un marco muestral adecuado para asegurar que las conclusiones a las que se arribe a partir del análisis de una *muestra* se puedan generalizar a toda la población. Las diferencias entre el marco muestral y la población de interés crea un *sesgo de cobertura*. Ante la presencia de dicho *sesgo*, los resultados de la *muestra* no tienen *validez externa* para toda la *población de interés*.

Método cuasi experimental. Métodos de *evaluación de impacto* que no dependen de la *asignación aleatoria* del tratamiento. Las *diferencias en diferencias,* el *diseño de regresión discontinua* y el *pareamiento* son ejemplos de métodos cuasi experimentales.

Método de control sintético. Un método de pareamiento específico que permite estimar el impacto en contextos donde una única *unidad* (como un país, una empresa o un hospital) es objeto de una *intervención* o es expuesto a un suceso. En lugar de comparar esta unidad tratada con un grupo de unidades no tratadas, el método utiliza información sobre las características de la unidad tratada y las unidades no tratadas para construir una unidad de comparación sintética o artificial, ponderando cada unidad no tratada de tal manera que la unidad de comparación sintética se parezca todo lo posible a la unidad tratada. Esto requiere una larga serie de observaciones a lo largo del tiempo, tanto de las características de la unidad tratada como de las unidades no tratadas. Esta combinación de unidades de comparación en una unidad sintética proporciona una mejor comparación para la unidad tratada que cualquier unidad no tratada individualmente.

Métodos mixtos. Enfoque analítico que combina datos cuantitativos y cualitativos.

Minería de datos. Práctica de manipular los datos en busca de resultados concretos.

Monitoreo. Proceso continuo de recopilar y analizar información para evaluar el desempeño de un proyecto, un programa o una política. El monitoreo suele hacer un seguimiento de los *insumos, actividades* y *productos*, aunque ocasionalmente también incluye los *resultados*. Se utiliza para fundamentar la gestión y las decisiones diarias. También se puede emplear para hacer un seguimiento del desempeño en relación con los resultados previstos, establecer comparaciones entre programas y analizar las tendencias a lo largo del tiempo.

Muestra aleatoria. Muestra extraída a partir de un *muestreo probabilístico*, por lo cual cada unidad en el *marco muestral* tiene una probabilidad conocida de ser extraída. Seleccionar una muestra aleatoria es la mejor manera de evitar una *muestra*

no representativa. El muestreo aleatorio no debería confundirse con la *asignación aleatoria*.

Muestra conglomerada. Una *muestra* compuesta de *clusters*.

Muestra estratificada. Se obtiene dividiendo la población de interés (*marco muestral*) en grupos (por ejemplo, hombres y mujeres) y luego definiendo una *muestra aleatoria* en cada grupo. Una muestra estratificada es una muestra probabilística: todas las *unidades* de cada grupo (o estrato) tienen la misma probabilidad de ser asignadas. Siempre que todos los grupos sean lo bastante grandes, el muestreo estratificado permite elaborar inferencias a propósito de los resultados no solo a nivel de la población sino también dentro de cada grupo.

Muestra. En Estadística, una muestra es un subconjunto de una *población de interés*. Normalmente, la población es muy grande, lo cual hace impracticable o imposible realizar un *censo* o un registro completo de todos sus valores. En cambio, los investigadores pueden seleccionar un subconjunto representativo de la población (utilizando un *marco muestral*) y recopilar estadísticas sobre la muestra. Estas se pueden utilizar para hacer inferencias o para extrapolar a la población. Este proceso se conoce como *muestreo*. Compárese con *censo*.

Muestreo. Proceso por el cual las unidades se extraen del *marco muestral* creado a partir de la *población de interés*. Se pueden utilizar diversas alternativas de procedimientos de muestreo. Los métodos de muestreo probabilístico son los más rigurosos, ya que asignan una probabilidad bien definida a cada unidad que será extraída. El *muestreo aleatorio*, el *muestreo aleatorio estratificado* y el *muestreo conglomerado* son métodos de muestreo probabilístico. El muestreo no probabilístico (por ejemplo, el muestreo intencional o por conveniencia) puede generar errores de muestreo.

Muestreo probabilístico. Proceso de muestreo que asigna una probabilidad bien definida a cada *unidad* que será extraída de un *marco muestral*. Incluye el *muestreo aleatorio*, el *muestreo aleatorio estratificado* y el *muestreo de* clusters.

Pareamiento por puntajes de propensión. Método de *pareamiento* que depende de los *puntajes de propensión* para encontrar el mejor *grupo de comparación* posible para un determinado *grupo de tratamiento*.

Pareamiento. Método no experimental de *evaluación de impacto* que utiliza grandes bases de datos y técnicas estadísticas para construir el mejor *grupo de comparación* posible para un determinado *grupo de tratamiento* sobre la base de características observables.

Población de interés. Grupo exhaustivo de todas las *unidades* (como individuos, hogares, empresas, centros) elegibles para recibir una intervención o un tratamiento, y para los cuales una *evaluación de impacto* se propone estimar los *impactos* del programa.

Potencia (o *potencia estadística*). Probabilidad de que una evaluación de impacto detecte un impacto (es decir, una diferencia entre el *grupo de tratamiento* y el *grupo de comparación*) cuando, de hecho, hay un impacto. La potencia es igual a 1 menos la probabilidad de un *error de tipo II*, que oscila entre 0 y 1. Los niveles habituales de

potencia son 0,8 y 0,9. Los niveles altos de potencia son más conservadores, lo que significa que hay una baja probabilidad de no detectar los impactos reales del programa.

Potencia estadística. La *potencia* de una prueba estadística es la probabilidad de que la prueba rechace la *hipótesis nula* cuando la *hipótesis alternativa* es verdadera (es decir, que no se cometerá un *error de tipo II*). A medida que la potencia aumenta, la probabilidad de un error de tipo II disminuye. La probabilidad de un error de tipo II se denomina tasa negativa falsa (β). Por lo tanto, la potencia es igual a $1 - \beta$.

Producto. Productos, bienes y servicios tangibles producidos (suministrados) directamente por las *actividades* de un programa. La generación de productos está directamente bajo el control del organismo ejecutor del programa. El uso de los productos por parte de los beneficiarios contribuye a cambios en los *resultados*.

Promoción aleatoria. Método de *variables instrumentales* para estimar los impactos de un programa. El método asigna de forma aleatoria a un subgrupo de *unidades* una *promoción*, o incentivo, para participar en el programa. La promoción aleatoria busca aumentar la participación voluntaria en un programa en una submuestra de la población seleccionada aleatoriamente. La promoción puede adoptar la forma de un incentivo, estímulo o información adicional que motiva a las unidades a inscribirse en el programa, sin influir directamente en el resultado de interés. De esta manera, el programa puede quedar abierto a todas las *unidades* elegibles.

Prueba de placebo. Prueba falsificada que se utiliza para evaluar si los supuestos de un método se mantienen. Por ejemplo, cuando se aplica el método de *diferencias en diferencias*, se puede implementar una prueba de placebo utilizando un grupo de tratamiento falso o un resultado falso, es decir: un grupo o resultado que se sabe que no se ve afectado por el programa. Las pruebas de placebo no pueden confirmar que los supuestos sean válidos, pero pueden poner de manifiesto los casos en que los supuestos no se sostienen.

Prueba de significancia. Prueba de si la *hipótesis alternativa* alcanza el nivel predeterminado de *significancia* con el fin de que esta se acepte de preferencia a la *hipótesis nula*. Si una prueba de significancia da un valor *p* menor que el nivel de *significancia* estadística (α), la *hipótesis nula* es rechazada.

Puntaje de propensión. En el contexto de la *evaluación de impacto,* el puntaje de propensión es la probabilidad de que una *unidad* participe en el programa sobre la base de las características observables. Esta puntuación es un número real entre 0 y 1 que resume la influencia de todas las características observables en la probabilidad de inscribirse en el programa.

Resultado. Resultado de interés que se mide a nivel de los beneficiarios del programa. Resultados que deben alcanzarse una vez que la población beneficiaria utilice los productos del proyecto. Los resultados no están directamente bajo el control de un organismo ejecutor del programa. En ellos influye tanto la implementación de un programa (las *actividades* y *productos* que genera) como las respuestas de las conductas de los beneficiarios expuestos a ese programa (el uso que los beneficiarios hacen de los beneficios a los que están expuestos). Un resultado puede ser intermedio o

final (de largo plazo). Los resultados finales son resultados más distantes. La distancia se puede interpretar en términos de tiempo (se tarda más en conseguir el resultado) o en términos de causalidad (se requieren numerosos vínculos causales para alcanzar el resultado y en ello influyen múltiples factores).

Selección. Se produce cuando la participación en el programa se basa en las preferencias, decisiones o características no observables de los participantes o de los administradores del programa.

Sesgo. En la *evaluación de impacto*, el sesgo es la diferencia entre el impacto que se calcula y el verdadero impacto del programa.

Sesgo de cobertura. Se produce cuando un *marco muestral* no coincide exactamente con la *población de interés*.

Sesgo de selección. El *impacto* estimado sufre un sesgo de selección cuando se desvía del *impacto* verdadero en presencia de la *selección*. Esto suele ocurrir cuando se correlacionan motivos no observados para participar en el programa con los *resultados*. Este sesgo normalmente acontece cuando el *grupo de comparación* es no elegible o se autoexcluye del tratamiento.

Sesgo de sustitución. Efecto no intencionado de la conducta que afecta al *grupo de comparación*. Las unidades que no fueron seleccionadas para recibir el programa pueden encontrar buenos sustitutos para el tratamiento a través de su propia iniciativa.

Significancia. La significancia estadística señala la probabilidad de cometer un *error de tipo I*; es decir, la probabilidad de detectar un impacto que en realidad no existe. El nivel de significancia suele señalarse con el símbolo griego α (alfa). Los niveles más habituales de significancia son del 10%, 5% y 1%. Cuanto menor sea el nivel de significancia, mayor será la confianza de que el impacto estimado es real. Por ejemplo, si el nivel de significancia se fija en 5%, se puede tener un 95% de confianza al concluir que el programa ha tenido impacto, si de hecho se observa un impacto significativo.

Simulaciones ex ante. *Evaluaciones* que utilizan datos disponibles para simular los efectos previstos de un programa o de la reforma de una política en los *resultados* de interés.

Supuesto de estabilidad del valor de la unidad de tratamiento (SUTVA). Requisito básico de que el *resultado* de una *unidad* no debería verse afectado por la asignación del tratamiento a otras unidades. Esto es necesario para asegurar que la *asignación aleatoria* produzca estimaciones de *impacto* no sesgadas.

Tamaño del efecto. Magnitud del cambio en un *resultado*, que es causado por una *intervención*.

Teoría del cambio. Explica los canales a través de los cuales los programas pueden influir en los *resultados* finales. Describe la lógica causal de cómo y por qué un programa, una modalidad de programa o una innovación de diseño en particular logrará sus resultados deseados. Una teoría del cambio es una pieza clave en cualquier *evaluación de impacto*, dada la focalización de causa y efecto de la investigación.

Tratamiento. Véase *intervención*.

Tratamiento en los tratados (TOT, por sus siglas en inglés). Las estimaciones TOT miden la diferencia en los *resultados* entre las *unidades* que en efecto reciben el tratamiento y el *grupo de comparación*.

Unidad. Persona, hogar, comunidad, empresa, escuela, hospital u otra unidad de observación que pueda ser objeto de un programa o verse afectada por él.

Validez externa. Una *evaluación* es externamente válida si la *muestra* de la evaluación representa con precisión a la población de *unidades* elegibles. Los resultados de la *evaluación* luego se pueden generalizar a la población de *unidades* elegibles. Estadísticamente, para que una *evaluación de impacto* sea externamente válida, la *muestra* de la evaluación debe ser representativa de la *población de interés*. Véase también *validez interna*.

Validez interna. Una *evaluación* es internamente válida si proporciona una estimación precisa del *contrafactual* mediante un *grupo de comparación* válido.

Variable. En la terminología estadística, se trata de un símbolo que representa un valor que puede variar.

Variable dependiente. Normalmente, es la variable de *resultado*. Se trata de la variable que hay que explicar, por oposición a las *variables explicativas*.

Variable explicativa. También conocida como variable "independiente". Se trata de una *variable* utilizada en el lado derecho de una regresión para ayudar a explicar la *variable dependiente* en el lado izquierdo de la regresión.

Variable instrumental (VI). También conocida como instrumento. Se basa en el uso de una fuente externa de variación para determinar la probabilidad de participación en el programa cuando la participación en el mismo está relacionada con los resultados potenciales. El instrumento se encuentra fuera del control de los participantes y no tiene relación con las características de los mismos.

Variables no observadas. Se trata de características no observables. Pueden incluir particularidades como la motivación, las preferencias u otros rasgos de la personalidad que son difíciles de medir.

ECO-AUDIT

Declaración de beneficios ambientales

El Grupo del Banco Mundial está comprometido a reducir su huella ambiental. En apoyo de este compromiso, la División de Publicaciones y Conocimiento impulsa las opciones de publicación electrónica y la tecnología de impresión bajo demanda, que funciona en centros regionales de todo el mundo. De forma conjunta, estas iniciativas permiten disminuir la cantidad de material impreso y acortar las distancias de envío, lo cual reduce el consumo de papel, el uso de químicos, las emisiones de gases de efecto invernadero y la basura. La División de Publicaciones y Conocimiento sigue las recomendaciones estándares para el uso de papel establecidas por la Green Press Initiative. La mayoría de nuestros libros se imprimen en papel certificado Forest Stewardship Council (FSC), que contiene cerca de un 50%-100% de material reciclado. Las fibras recicladas del papel de nuestros libros no se blanquean o bien se las blanquea mediante un proceso totalmente libre de cloro, o con cloro elemental o mejorado. Para más información sobre la filosofía ambientalista del Banco, visítese el sitio http://www.worldbank.org/corporateresponsibility.

green press INITIATIVE

"*La evaluación de impacto en la práctica* es simplemente una joya. Propone un enfoque de la evaluación de impacto que busca ser creíble científicamente y, al mismo tiempo, reconoce las realidades prácticas de realizar este tipo de trabajo en el campo. A lo largo de todo el libro hay insumos valiosos en estas dos dimensiones. Yo asigno todo el tiempo este libro como material de lectura a la hora de capacitar profesionales interesados en la realización, la puesta en marcha, o el consumo de evaluaciones de impacto."

—**Dan Levy,** *Catedrático Senior de Políticas Públicas y Director de la Iniciativa para el Fortalecimiento del Aprendizaje y la Excelencia de la Enseñanza*

"*La evaluación de impacto en la práctica* es una gran contribución a la agenda de desarrollo contemporánea. Es un recurso de gran valor para los evaluadores de los gobiernos y organismos de desarrollo, así como en las universidades y centros de investigación."

—**Leonard Wantchekon,** *Profesor de Política y Relaciones Internacionales, Universidad de Princeton; Fundador y Presidente de la Escuela Africana de Economía*

"El propósito de este libro es ofrecer una guía accesible, comprehensiva y clara sobre las evaluaciones de impacto. El material, que va desde la motivación de la evaluación de impacto hasta las ventajas de las diferentes metodologías, cálculos de potencia y costos, se explica muy claramente, y la cobertura es impresionante. Este libro se convertirá en una guía muy consultada y utilizada que afectará la formulación de políticas durante los próximos años."

—**Orazio Attanasio,** *Profesor de Economía, University College of London; Director del Centro de Evaluación y Políticas de Desarrollo, Instituto de Estudios Fiscales, Reino Unido*

"La versión actualizada de este libro extraordinario llega en un momento crítico: la cultura y el interés por la evaluación están creciendo y necesitan el apoyo de un trabajo técnico de calidad. *La evaluación de impacto en la práctica* es un recurso esencial para evaluadores, programas sociales, ministerios, y todos aquellos comprometidos con la toma de decisiones con base en buena evidencia. Esta obra es cada vez más importante a medida que la comunidad de desarrollo global trabaja para reducir la pobreza y alcanzar la Agenda de Desarrollo Sostenible 2030."

—**Gonzalo Hernández,** *Secretario Ejecutivo, Consejo Nacional de Evaluación para la Política de Desarrollo Social, México*

El material adicional de *La evaluación de impacto en la práctica* se encuentra disponible en el sitio web: http://www.worldbank.org/ieinpractice.

Strategic Impact Evaluation Fund

BID
Banco Interamericano de Desarrollo

GRUPO BANCO MUNDIAL

ISBN 978-1-4648-0888-3

90000
9 781464 808883

SKU 210888