

Evaluación de Programas Contra la Pobreza

Martin Ravallion¹

Grupo de Investigación sobre el Desarrollo, Banco Mundial

Resumen: Este capítulo ofrece una visión crítica de los métodos disponibles para el análisis contrafactual *ex post* de los programas que se asignan exclusivamente a individuos, hogares o localidades. El debate abarca métodos experimentales y no experimentales (entre ellos, correspondencia del puntaje de propensión [*propensity-score matching*], diseños de discontinuidad, doble y triple diferencia y variables instrumentales). Al analizar los problemas surgidos en la aplicación de los métodos en los programas de lucha contra la pobreza en países en desarrollo, surgen dos consecuencias importantes. La primera es que, a pesar de lo que digan sus defensores, ningún método predomina sobre otro. Las evaluaciones rigurosas de políticas deben mantener una actitud abierta respecto a la metodología y adaptarse a las limitaciones de los datos, el problema y el entorno. La segunda es que, si se desean obtener consecuencias útiles de las evaluaciones en el futuro, será necesario recurrir a datos y métodos más relevantes que la clásica evaluación de impacto tipo “caja negra”, basada en resultados promedio.

Contenidos

1.	Introducción	2
2.	El problema de la evaluación arquetípica	3
3.	Cuestiones generales	8
4.	Experimentos sociales	19
5.	Métodos basados en el puntaje de propensión (<i>propensity-score</i>)	26
6.	Cómo aprovechar el diseño del programa	35
7.	Diferencias de orden superior	39
8.	Flexibilización de la exogeneidad condicional	50
9.	Conocimientos que se obtienen de las evaluaciones	61

¹ Estas son opiniones del autor y no deben atribuirse al Banco Mundial ni a ninguna de sus organizaciones afiliadas. El autor agradece por sus comentarios a Pedro Carneiro, Aline Coudouel, Jishnu Das, Jed Friedman, Emanuela Galasso, Markus Goldstein, Jose Garcia-Montalvo, David McKenzie, Alice Mesnard, Ren Mu, Norbert Schady, Paul Schultz, Emmanuel Skoufias, Petra Todd, Dominique van de Walle, y a los participantes de varias presentaciones del Banco Mundial y de un taller del autor realizado en el Rockefeller Foundation Center de Bellagio, en Italia, en mayo de 2005.

10. Conclusiones	74
Figuras	76
Referencias	79

1. Introducción

Los gobiernos, donantes y la comunidad de desarrollo en su conjunto solicitan cada vez más pruebas concretas sobre el impacto de los programas públicos que afirman reducir la pobreza. ¿Sabemos si tales intervenciones realmente funcionan? ¿Cuán efectivas son? Las antiguas “evaluaciones”, que sólo proporcionaban datos cualitativos sobre los procesos y no evaluaban los resultados relacionados con contrafácticos explícitos y relevantes a las políticas, son consideradas deficientes en la actualidad.

Este capítulo ofrece una revisión crítica de los principales métodos disponibles para el análisis contrafactual de los programas que son asignados exclusivamente a ciertas unidades observacionales. Estas unidades pueden ser personas, hogares, poblaciones o zonas geográficas más extensas. La característica principal es que algunas unidades se benefician con el programa y otras no. Por ejemplo, un fondo social puede solicitar a las comunidades que presenten propuestas, especialmente a aquellas que viven en zonas más pobres. No obstante, algunas zonas no son incluidas, mientras que otras sí pero son rechazadas.² O un programa para desempleados (que exige que los beneficiarios de la asistencia social trabajen en contraprestación por los beneficios que obtienen) supone ingresos adicionales para los trabajadores que participan del programa y ganancias para los residentes de las zonas en las que se realizan los trabajos, mientras que otras personas no reciben ayuda alguna. O transferencias de dinero en efectivo destinadas exclusivamente a hogares considerados elegibles según un determinado criterio.

Luego de la perspectiva general de la formulación arquetípica del problema de evaluación en la literatura referida al tema, la mayor parte del capítulo examina los métodos principales que se utilizan en la práctica. El debate analiza los supuestos en los que se basa cada método para

² Los fondos sociales proporcionan apoyo financiero para una amplia variedad de proyectos comunitarios, con énfasis especial en la participación de los habitantes locales en la propuesta e implementación de los proyectos específicos.

identificar el impacto de un programa, la manera en que se comparan los métodos entre sí y la información disponible sobre sus resultados. Los ejemplos se extraen principalmente de evaluaciones realizadas en países en desarrollo. La penúltima sección intenta ir un poco más allá: indaga sobre el modo en que evaluaciones futuras pueden ser más útiles para la obtención de información y la formulación de políticas. La sección final sugiere dos consecuencias importantes de este estudio.

2. El problema de la evaluación arquetípica

Una evaluación de impacto tiene como objetivo evaluar los resultados de un programa comparándolo con un contrafáctico explícito; por ejemplo, cómo sería la situación en ausencia del programa. Cuando el programa se encuentra en funcionamiento, la tarea de evaluación es *ex-post*. (Esto incluye la evaluación de un proyecto piloto como entrada para la evaluación *ex-ante*, a fin de determinar si el proyecto debe ampliarse o no). Sin embargo, la realización de una evaluación *ex-post* no significa que ésta deba comenzar una vez que finalice el programa, o incluso después de que éste haya comenzado. Las mejores evaluaciones *ex-post* se diseñan e implementan *ex-ante*, con frecuencia mientras se desarrolla el programa.

En primer lugar es necesario establecer cuál será el indicador de resultado observable más relevante para los objetivos del programa. Digamos que este indicador es una variable aleatoria, Y , con una media de población $E(Y)$. En los programas de lucha contra la pobreza, por lo general, el objetivo se define en términos de los ingresos o gastos (sobre el consumo) del hogar normalizado por una línea de pobreza específica del hogar (que refleja las diferencias en los precios y en el tamaño y composición del grupo familiar). Si deseamos conocer el impacto del programa de lucha contra la pobreza, podemos establecer $Y=1$ para los “pobres” versus $Y=0$

para los “no pobres”, de manera tal que $E(Y)$ sea el índice de recuento de la pobreza.³ Con frecuencia se necesita más de un indicador. Considere, por ejemplo, un esquema por el que realizan transferencias de dinero a familias pobres con la sola condición de que los padres efectúen inversiones de recursos humanos en sus hijos.⁴ Los resultados relevantes deben incluir, por supuesto, una medición del nivel de pobreza actual pero, para este tipo de programa, también es necesario evaluar el nivel de escolaridad y el estado de salud de los hijos, ya que estos datos se pueden interpretar como indicadores de pobreza futura.

Suponemos que nuestros datos incluyen una observación de Y_i para cada unidad i en una muestra de tamaño n . Algunas unidades reciben el programa, en cuyo caso se consideran “tratadas,” y dejamos $T_i = 1$, mientras que para las unidades “no tratadas” utilizamos $T_i = 0$.⁵ La formulación arquetípica del problema de evaluación sigue el modelo de Rubin (1974) al postular dos resultados posibles para cada i ; el valor de Y_i bajo tratamiento se indica en Y_i^T mientras que según el contrafáctico explícito de no recibir tratamiento es Y_i^C .⁶ La unidad i obtiene una ganancia: $G_i \equiv Y_i^T - Y_i^C$. En la literatura, G_i se denomina “ganancia”, “impacto” o “efecto causal” del programa para la unidad i .

Siguiendo la línea propuesta por la mayor parte de la literatura, este capítulo se concentrará principalmente en la estimación de los impactos medios (aunque se señalarán

³ Transformar la información de niveles de vida en una variable binaria no es necesariamente el enfoque más eficiente para medir el impacto de la pobreza. Volveremos a este punto más adelante.

⁴ Al parecer, el primer programa de este tipo implementado en un país en desarrollo fue el programa *Food-for-Education* [Alimentos por educación], ahora denominado *Cash-for-Education* [Dinero por educación] introducido por el gobierno de Bangladesh en 1993. Un ejemplo famoso de este tipo de programa es el Programa de Educación, Salud y Alimentación (PROGRESA) (ahora denominado “*Oportunidades*”), implementado por el gobierno de México en 1997.

⁵ Las connotaciones biomédicas de la palabra “tratamiento” son poco acertadas en el contexto de política social, pero el uso casi universal de este término en la literatura de evaluación hace que sea difícil evitarlo.

⁶ En la literatura, Y_1 o $Y(1)$ y Y_0 o $Y(0)$ se utilizan comúnmente para Y^T y Y^C . Mi notación (según el modelo de Holland, 1986) facilita la identificación de los grupos, especialmente más adelante cuando introduzco subíndices de tiempo.

implicaciones para otros parámetros de impacto a medida que vayan surgiendo). La medición de impacto medio más utilizada es el efecto medio del tratamiento sobre los tratados:

$TT \equiv E(G|T = 1)$. En el contexto de un programa de lucha contra la pobreza, TT es el impacto medio sobre la pobreza entre los individuos que reciben el programa. Otro parámetro de interés es el efecto medio del tratamiento sobre los no tratados, $TU \equiv E(G|T = 0)$, y el efecto medio del tratamiento combinado (ATE):

$$ATE \equiv E(G) = TT \Pr(T = 1) + TU \Pr(T = 0)$$

(Cada uno de estos parámetros posee una estimación de muestra correspondiente). Con frecuencia deseamos conocer los impactos medios condicionales, $TT(X) \equiv E(G|X, T = 1)$, $TU(X) \equiv E(G|X, T = 0)$ y $ATE(X) \equiv E(G|X)$, para un vector de covariantes X (incluida la unidad como un elemento). El método más común para introducir X supone que los resultados son lineales en sus parámetros y los términos de error (μ^T y μ^C), quedando definido como:

$$Y_i^T = X_i \beta^T + \mu_i^T \quad (i=1, \dots, n) \quad (1.1)$$

$$Y_i^C = X_i \beta^C + \mu_i^C \quad (i=1, \dots, n) \quad (1.2)$$

Definimos los parámetros β^T y β^C de manera tal que X sea exógena

($E(\mu^T|X) = E(\mu^C|X) = 0$).⁷ Los impactos medios condicionales son entonces:

$$TT(X) = ATE(X) + E(\mu^T - \mu^C|X, T = 1)$$

$$TU(X) = ATE(X) + E(\mu^T - \mu^C|X, T = 0)$$

$$ATE(X) = X(\beta^T - \beta^C)$$

¿Cómo podemos calcular estos parámetros de impacto a partir de los datos disponibles? La literatura reconoce desde hace tiempo que la evaluación de impacto es

⁷ Esto es posible debido a que no es necesario aislar los efectos directos de X de aquellos que operan mediante variables omitidas correlacionadas con X .

esencialmente un problema de falta de datos, ya que es físicamente imposible medir resultados en una misma persona en dos condiciones diferentes al mismo tiempo (participando y no participando en un programa). Se supone que podemos observar T_i , Y_i^T para $T_i = 1$, Y_i^C para $T_i = 0$, y (en consecuencia) $Y_i = T_i Y_i^T + (1 - T_i) Y_i^C$. Pero entonces G_i no es directamente observable para ninguna i ya que faltan datos en Y_i^T para $T_i = 0$ y en Y_i^C para $T_i = 1$. Tampoco es posible identificar los impactos medios sin realizar suposiciones adicionales; ni $E(Y^C|T = 1)$ (según se requiere para calcular TT y ATE) ni $E(Y^T|T = 0)$ (según se requiere para calcular TU y ATE) se pueden calcular de manera directa a partir de los datos. Las ecuaciones (1.1) y (1.2) tampoco constituyen un modelo válido, debido a los datos que faltan.

Con los datos que probablemente estén disponibles, un lugar obvio donde comenzar es la diferencia simple (D) de los resultados medios entre los participantes y los no participantes.

$$D(X) \equiv E[Y^T|X, T = 1] - E[Y^C|X, T = 0] \quad (2)$$

Este valor se puede calcular por la diferencia en los promedios de las muestras correspondientes o (de manera equivalente) por el coeficiente de regresión del método de mínimos cuadrados ordinarios (*Ordinary Least Squares* u *OLS*, por su sigla en inglés) de Y en T . Para el modelo paramétrico con controles, se estima (1.1) en la submuestra de participantes y (1.2) en el resto de la muestra, lo que arrojaría el siguiente resultado:

$$Y_i^T = X_i \beta^T + \mu_i^T \text{ si } T_i = 1 \quad (3.1)$$

$$Y_i^C = X_i \beta^C + \mu_i^C \text{ si } T_i = 0 \quad (3.2)$$

De manera similar, se puede utilizar la práctica más común en el trabajo aplicado que consiste en calcular una regresión simple (“switching”) para la medición de resultados observada en la muestra combinada, lo que proporciona una especificación de “coeficientes aleatorios”:⁸

$$Y_i = X_i\beta^C + X_i(\beta^T - \beta^C)T_i + \varepsilon_i \quad (i=1, \dots, n) \quad (4)$$

donde $\varepsilon_i = T_i(\mu_i^T - \mu_i^C) + \mu_i^C$. En la práctica, un conocido caso especial es el modelo de impacto común, que supone $G_i = ATE = TT = TU$ para todas las i , de modo que (4) se reduce a:

$$Y_i = ATE.T_i + X_i\beta^C + \mu_i^C \quad (5)$$

Un modelo menos restrictivo sólo impone la condición de que los efectos latentes sean los mismos para los dos grupos ($\mu_i^T = \mu_i^C$), de modo que los efectos de interacción con X se mantengan. Este modelo es a veces denominado modelo de efectos comunes.⁹

Si bien todos éstos son puntos de partida razonables para una evaluación y cuentan con un obvio interés descriptivo, es necesario contar con más supuestos para asegurar estimaciones imparciales de los parámetros de impacto. Para entender por qué, considere la diferencia en los resultados medios entre participantes y no participantes (ecuación 2). Esto puede expresarse de la siguiente manera:

$$D(X) = TT(X) + B^{TT}(X) \quad (6)$$

donde:¹⁰

$$B^{TT}(X) \equiv E[Y^C | X, T = 1] - E[Y^C | X, T = 0] \quad (7)$$

⁸ La ecuación (4) surge de (3.1) y (3.2) utilizando la identidad $Y_i = T_i Y_i^T + (1 - T_i) Y_i^C$.

⁹ La justificación para estas especializaciones de (4) es rara vez obvia; se debe presumir la heterogeneidad en los impactos a menos que exista evidencia concreta sobre lo contrario. Volveré a este punto más adelante.

¹⁰ De manera similar $B^{TU}(X) \equiv E(Y^T | X, T = 1) - E(Y^T | X, T = 0)$ y

$B^{ATE}(X) = B^{TT}(X) \Pr(T = 1) + B^{TU}(X) \Pr(T = 0)$ en notación obvia.

es el sesgo al utilizar $D(X)$ para calcular $TT(X)$; B^{TT} se denomina sesgo de selección en gran parte de la literatura sobre evaluación. Claramente, la diferencia en los valores medios (o coeficiente de regresión OLS en T) sólo proporciona el efecto medio del tratamiento sobre los tratados si los resultados medios contrafactuales no varían con el tratamiento, es decir, $B^{TT} = 0$. En términos del modelo paramétrico anteriormente descrito, esto equivale a suponer que $E[\mu^C | X, T = 1] = E[\mu^C | X, T = 0] = 0$, lo cual asegura que OLS proporciona estimaciones coherentes de (5). Si esto también se aplica para μ^T , OLS proporcionará estimaciones coherentes de (4). Haré referencia al supuesto $E(\mu^C | X, T = t) = E(\mu^T | X, T = t) = 0$ para $t=0,1$ como “exogeneidad condicional de la implementación del programa”.¹¹ El resto de este capítulo se organiza en torno a los principales métodos que se utilizan en la práctica para calcular los impactos del programa en la formulación arquetípica del problema de evaluación antes mencionado. Una manera obvia de asegurarse que $B^{TT} = 0$ es implementar el programa de manera aleatoria condicional a X . De esta manera se trabaja con una evaluación experimental, que se tratará con mayor profundidad en la sección 4. En contraposición a lo anterior, en una evaluación no experimental (NX) (también llamada “estudio observacional” o “evaluación cuasi-experimental”) el programa no se implementa de manera aleatoria.¹² La mayor parte del capítulo trata de los métodos NX. Estos dos tipos de evaluaciones difieren en los supuestos en los que se basan para identificar impactos. Los métodos principales se dividen en dos grandes grupos, según el supuesto de identificación (no anidado) que utilizan. El primer grupo supone exogeneidad condicional de implementación, o el supuesto un tanto más débil de exogeneidad para cambios en la implementación en relación con los cambios en los resultados. Las secciones 5 y 6 investigan los métodos de diferencia simple que comparan resultados entre muestras de participantes y no participantes (posiblemente seleccionadas cuidadosamente). La sección 7 se centra en los métodos de doble o triple diferencia. Estos métodos utilizan datos sobre cambios en resultados e implementación, tales como los que se detectan al observar los resultados para ambos grupos antes y después del comienzo del programa.

El segundo conjunto de métodos no supone exogeneidad condicional (ni en diferencia simple ni de orden superior). El principal supuesto alternativo encontrado en los trabajos aplicados es la existencia de una variable instrumental que no altera los resultados según la participación (y otras covariantes de resultado) pero es, no obstante, una covariante de participación. La variable instrumental, por lo tanto, aísla una parte de la variación de la implementación del programa que puede tratarse como exógena. Este método se analiza en la sección 8, junto con otras alternativas (menos conocidas pero prometedoras).

¹¹ En la literatura sobre evaluación, este supuesto también se denomina “selección basada en características observables”, “asignación sin factores de confusión” o “asignación ignorable” (aunque estos dos últimos términos se refieren generalmente al supuesto más firme de que Y^T y Y^C son independientes de T dado X).

¹² Como veremos más adelante, en la práctica los métodos experimentales y NX en ocasiones se combinan, aunque la distinción resulta útil a fines expositivos.

Algunos evaluadores prefieren utilizar uno de estos dos supuestos de identificación antes que el otro. Sin embargo, no existe una base sólida *a priori* que justifique una preferencia fija en esta opción, que debe realizarse teniendo en cuenta las particularidades de cada caso y basándose en la información sobre el programa, su entorno y (fundamentalmente) los datos disponibles.

3. Cuestiones generales

Con frecuencia, el primer problema que surge en la práctica es que las principales partes interesadas estén de acuerdo en realizar una evaluación de impacto. Es posible que esto represente una amenaza para los intereses personales de algunas personas, incluido el personal del proyecto. También pueden presentarse objeciones éticas. La objeción más común a la realización de una evaluación de impacto es la que afirma que para que un grupo de comparación sea válido, debe incluir personas igualmente necesitadas entre los participantes, en cuyo caso la única opción éticamente aceptable es ayudarlos, en lugar de simplemente observarlos de manera pasiva con fines evaluativos. Al parecer, diferentes versiones de este argumento han impedido o retrasado muchas evaluaciones.

Las objeciones éticas en contra de las evaluaciones de impacto para los programas de lucha contra la pobreza deben tomarse seriamente. Claramente, las objeciones tienen mucho más peso si se ha negado la participación en el programa a personas elegibles con el fin de realizar la evaluación y la información obtenida a partir de la evaluación no los beneficia. Sin embargo, el motivo principal por el cual es posible armar grupos de comparación válidos, es que normalmente los recursos fiscales no son suficientes para cubrir las necesidades de todas las personas. Si bien se podría objetar este hecho, no es una objeción contra la evaluación *per se*. Además, el conocimiento sobre el impacto de los programas puede tener gran relevancia en los recursos disponibles para luchar contra la pobreza. Las personas pobres se benefician con la

realización de buenas evaluaciones, ya que se pueden eliminar programas ineficientes de lucha contra la pobreza e identificar aquellos realmente útiles.

Una vez conseguida la aceptación para realizar la evaluación, se deberán resolver tres tipos de problemas. El primero es la selección no aleatoria; el segundo, la existencia de efectos derivados, lo que hace difícil ubicar los impactos de un programa sólo entre sus participantes directos. Después de examinar estas cuestiones, la sección analiza un tercer conjunto de problemas generales relacionados con datos y medición.

¿Existe sesgo de selección? La asignación de un programa de lucha contra la pobreza por lo general involucra una implementación específica dirigida, que refleja tanto las selecciones realizadas por las personas elegibles como la asignación administrativa de las oportunidades de participar. Esto no es un problema si las X de los datos capturan los determinantes “no ignorables” de la implementación; es decir, aquellos que guardan relación con los resultados. No obstante, todo factor no ignorable latente –no observado por el evaluador pero conocido por las personas que deciden la participación y con influencia en los resultados– provocará un sesgo en el estimador de impacto sobre la base de las diferencias de promedios entre los participantes y los no participantes, o de cualesquiera de los métodos de regresión paramétrica viables. El análisis que sigue comienza con los sesgos de selección surgidos de controles inadecuados de la heterogeneidad observable y luego se centra en sesgos surgidos de características no observables.

Un tema importante en cualquier evaluación NX es saber si el proceso de selección del programa a evaluar es capturado adecuadamente por las variables de control X . Este tema no puede separarse estrictamente del problema de la asignación no aleatoria condicional a la presencia de características observables. Por supuesto, no es posible determinar si la exogeneidad condicional de la implementación es un supuesto admisible, sin antes

establecer un supuesto que haya resuelto correctamente la heterogeneidad observable, a pesar de las variables condicionantes.

Un aspecto problemático de los métodos tradicionales de regresión lineal es que las ecuaciones (3) y (4) se ocupan de la selección de características observables de un modo bastante especial, porque los controles se introducen de una manera lineal en sus parámetros. Esta suposición *ad hoc* rara vez se justifica, excepto por una cuestión de comodidad computacional (con poco fundamento hoy en día). La sección 5 examinará los métodos no paramétricos que procuran tratar esta fuente de sesgos de una manera más general.

En las evaluaciones NX de los programas de lucha contra la pobreza a veces puede resultar difícil garantizar que las características observables se equilibren entre las observaciones de tratamiento y las de comparación. Cuando la implementación de un programa es independiente de los resultados dadas las características observables (lo que implica exogeneidad condicional, según se define en la sección 2) el dato estadístico relevante a sopesar entre los dos grupos es la probabilidad condicional de participación en el programa, denominada “*propensity score*” (Rosenbaum y Rubin, 1983).¹³ La región de probabilidades donde se puede encontrar un grupo de comparación válido se denomina región de soporte común (*region of common support*), como se muestra en la Figura 1.

Para ilustrar el problema potencial de soporte común al evaluar un programa de lucha contra la pobreza, supongamos que la implementación se determina mediante una calificación socioeconómica de la familia (“*proxy-means test*”), que generalmente se utiliza para orientar los programas de lucha contra la pobreza a sectores específicos de países en desarrollo. Con esta calificación se otorga un puntaje a todos los posibles participantes en función de las

¹³ El puntaje de propensión (*propensity score*) juega un papel importante en varios métodos NX, como veremos más adelante en la sección 5.

características observadas. Cuando se aplica de manera rigurosa, el programa se asigna únicamente si el puntaje de una unidad es inferior a un nivel crítico, determinado por la asignación presupuestaria del esquema. (El puntaje mínimo es no decreciente en el presupuesto en condiciones viables.) Con una participación del 100%, no existe un valor de puntaje que contemple a los participantes y a los no participantes en una muestra, sin importar del tamaño que sea. Éste es un ejemplo de lo que en ocasiones la literatura sobre evaluación denomina “falla del soporte común”. El problema es bastante simple: ¿cómo inferir el contrafáctico para participantes sobre la base de no participantes que no comparten las mismas características, según demuestran los puntajes de la prueba de calificación socioeconómica?

Claramente, es lógico que exista una seria preocupación sobre la validez de los diseños de grupos de comparación para identificar impactos.¹⁴ Si bien este ejemplo tiene valor pedagógico, es un caso extremo. Afortunadamente, en la práctica, con frecuencia existe algún grado de discrepancia en la aplicación de la prueba de calificación socioeconómica y, por lo general, existe una cobertura incompleta de las personas que obtienen el puntaje mínimo.

Normalmente, se tiene que truncar la muestra de no participantes para garantizar el soporte común y, fuera de la ineficiencia que supone recolectar datos innecesarios, esto no supone un problema. Más preocupante es que una submuestra no aleatoria de participantes tenga que eliminarse por falta de comparadores suficientemente similares. Esto sugiere una compensación entre dos fuentes de sesgo. Por un lado, existe la necesidad de garantizar la comparabilidad en términos de características iniciales. Por otro lado, esto crea un posible sesgo

¹⁴ Si no se necesita conocer el impacto para el grupo de tratamiento en su conjunto, el problema se reduce. Por ejemplo, considere la opción de aumentar la asignación presupuestaria del programa elevando el puntaje mínimo en la prueba de calificación socioeconómica. En este caso, sólo habría que centrarse en los impactos de un vecindario donde se registró la puntuación mínima. La sección 6 analiza con mayor detalle los “diseños de discontinuidad” para tales casos.

de muestra en interferencias sobre impacto, a tal grado que es posible que se tengan que eliminar unidades de tratamiento para lograr comparabilidad.

La participación no aleatoria también supone la posibilidad de sesgo si alguna de las variables que afectan los resultados y la implementación del programa no son observadas por el evaluador. En ese caso, no se pueden atribuir al programa las características observadas $D(X)$. Las diferencias en medios condicionales que vemos en los datos pueden deberse a que los participantes del programa fueron elegidos expresamente mediante un proceso no observado en su totalidad. El estimador de impacto tiene un grado de parcialidad o sesgo que se refleja en la ecuación (7). Cuando la participación en el programa depende del sujeto, debe existir una presunción razonable de que la selección para participar en el programa depende de las ganancias obtenidas por dicha participación, que no son observadas en su totalidad por el evaluador. Por ejemplo, supongamos que el proceso de selección latente discriminara a los pobres, es decir, $E[Y^C|X, T = 1] > E[Y^C|X, T = 0]$, donde Y es el ingreso relativo a la línea de pobreza. Entonces $D(X)$ sobrestimaré el impacto del programa. Un proceso de selección latente que favoreciera a los pobres tendrá el efecto contrario.

En términos de la formulación clásica paramétrica del problema de evaluación en la sección 2, si los participantes tienen atributos latentes que producen resultados más altos que los no participantes (según la X determinada), entonces los términos de error en la ecuación para los participantes (3.1) se centrará hacia la derecha en relación con los términos de los no participantes (3.2). El término de error en (4) no desaparecerá en la expectativa y el OLS proporcionará cálculos sesgados y contradictorios. (Nuevamente, la preocupación sobre esta fuente de sesgo no puede separarse de la pregunta referida al grado de control ejercido sobre la heterogeneidad observable).

Varios ejemplos de estudios de replicación sugieren que el sesgo de selección puede ser un problema serio en las estimaciones de impacto NX en casos específicos. Destacados estudios de Lalonde (1986) y de Fraker y Maynard (1987) encontraron importantes sesgos al comparar los resultados de varios métodos NX con evaluaciones aleatorias de un programa de capacitación de los Estados Unidos. (Varios métodos NX también dieron resultados bastante diferentes, aunque esto no es sorprendente si se tiene en cuenta que se basaron en supuestos distintos). De manera similar, Glewwe et ál. (2004) llegan a la conclusión de que los métodos NX muestran un mayor impacto estimado de “*flip charts*” en los puntajes de pruebas de niños en edad escolar de Kenia que los obtenidos por un experimento, y sostienen que las diferencias se deben a los sesgos presentes en los métodos NX. En un interesante meta-estudio, Glazerman et ál. (2003), se analizaron 12 estudios de replicación sobre el impacto de algunos programas de capacitación y empleo en los ingresos. Cada estudio comparó las estimaciones de impacto NX con los resultados de un experimento social del mismo programa. Se encontraron importantes discrepancias en algunos casos; se interpretó que estas discrepancias se debieron a sesgos en las estimaciones NX.

Utilizando un enfoque diferente para comprobar la eficacia de los métodos NX, van de Walle (2002) ofrece un ejemplo para la evaluación de rutas en zonas rurales. Una comparación simplista de los ingresos económicos de poblaciones que tienen rutas en zonas rurales en relación con poblaciones que no las tienen, arroja importantes aumentos en los ingresos cuando, en rigor, tales aumentos no existen. Van de Walle utilizó métodos de simulación en los que los datos se extraían de un modelo en el cual los verdaderos beneficios eran conocidos con certeza, y las rutas se ubicaban en parte como una función de los ingresos medios de las poblaciones. Una sobrestimación aparentemente pequeña de los ingresos de las

poblaciones en la determinación de la ubicación de rutas fue suficiente para sesgar seriamente el impacto medio estimado.

Por supuesto, no podemos rechazar el uso de métodos NX en otras aplicaciones sobre la base de estos estudios. Posiblemente la lección sea que es necesario contar con mejores datos y métodos, basados en el conocimiento anterior del funcionamiento de tales programas. Ante la presencia de graves problemas de datos no es de extrañar que los estudios observacionales no hayan sido de gran ayuda en la corrección de sesgos de selección. Por ejemplo, en un convincente análisis crítico del estudio de Lalonde, Heckman y Smith (1995) señalan (entre otras cosas) que los datos utilizados contenían muy poca información relevante a la elegibilidad para el programa estudiado; que los métodos utilizados tenían una capacidad limitada para resolver sesgos de selección y que el estudio no incluía pruebas de especificación adecuadas.¹⁵ Heckman y Hotz (1989) sostienen que las pruebas de especificación adecuadas pueden revelar los métodos NX problemáticos en el estudio de Lalonde, y que los métodos que superan sus pruebas proporcionan resultados cercanos a los del experimento social.

Los 12 estudios utilizados por Glazerman et ál. (2003) les proporcionaron más de 1100 observaciones de estimaciones de impacto pareadas; una experimental y una NX. Los autores hicieron una regresión de los sesgos estimados utilizando regresores que describían los métodos NX. Descubrieron que los métodos NX tenían mejores resultados (más cercanos al resultado experimental) cuando los grupos de comparación eran seleccionados cuidadosamente sobre la base de diferencias observables (utilizando métodos de regresión, pareo o una combinación de ambos). Sin embargo, también descubrieron que los métodos econométricos estándar utilizados para resolver el sesgo de selección debido a características no observables

¹⁵ Consulte también el debate descrito en Heckman et ál. (1999).

que usan una función de control y/o una variable instrumental, tendían a umentar la divergencia entre las dos estimaciones.

Estos hallazgos advierten contra la presunción de que los métodos NX más ambiciosos y aparentemente sofisticados tendrán mejores resultados en la reducción del sesgo total. La literatura también señala la importancia de las pruebas de especificación y el escrutinio crítico de los supuestos de cada estimador. Este capítulo volverá a tratar este punto en el contexto de estimadores específicos.

¿Existen impactos ocultos para los “no participantes”? La formulación clásica del problema de evaluación descrita en la sección 2 asume que podemos observar los resultados bajo tratamiento (Y_i^T) para los participantes ($T_i = 1$) y el resultado contrafactual (Y_i^C) para los no participantes ($T_i = 0$). Luego, podemos observar un grupo de comparación que no ha sido afectado por el programa en ningún aspecto. No obstante, esto puede ser un supuesto problemático para ciertos programas de lucha contra la pobreza. Supongamos que estamos evaluando un programa de ayuda económica laboral en el cual el gobierno se compromete a dar trabajo pagando un salario estipulado. Éste fue el objetivo del famoso *Employment Guarantee Scheme* (Esquema de empleo garantizado o EGS) del estado de Maharashtra, en India; en 2005 el gobierno nacional implementó una versión de este esquema en todo el país. La atracción de un EGS como red de seguridad se debe a que el acceso al programa es universal (cualquier persona que solicite ayuda puede obtenerla), pero todos los participantes deben trabajar para obtener beneficios y con un salario que se considera bajo en el contexto específico. La universalidad de acceso significa que el esquema puede proporcionar un seguro efectivo contra riesgos. Para quienes proponen este esquema, el requisito de trabajar por un salario bajo ayuda a que efectivamente sus beneficiarios sean los sectores de menores ingresos.

Se lo puede considerar un programa asignado por el hecho de que existe una clara distinción entre los “participantes” y los “no participantes”. Y a primera vista puede parecer apropiado recopilar datos de encuestas de ambos grupos y comparar indicadores de resultados entre ambos, como una forma de identificar impactos (posiblemente después de corregir cualquier heterogeneidad observable). Sin embargo, este diseño clásico de evaluación puede arrojar un resultado seriamente sesgado. Las ganancias derivadas de tales programas deben llegar al mercado de trabajo privado. Si la garantía de empleo es efectiva, el esquema establecerá un límite inferior firme para toda la distribución de salarios, suponiendo que ningún trabajador físicamente apto aceptaría un trabajo diferente al que ofrece EGS por un salario inferior al de EGS. Por lo tanto, incluso si se elige el grupo de comparación perfecto desde el punto de vista observacional, se puede concluir que el esquema no tiene impacto alguno, ya que los salarios serán los mismos para los participantes que para los no participantes. Pero eso pasaría totalmente por alto el impacto, que puede ser importante para ambos grupos.

Tales efectos derivados también pueden surgir por la conducta de los gobiernos. Con frecuencia, no queda claro si los recursos transferidos a los participantes realmente financiaron el proyecto identificado. Hasta cierto punto, toda ayuda externa es intercambiable. Sí, se puede determinar mediante supervisión si el subproyecto propuesto se completó realmente. Pero nadie puede descartar la posibilidad de que pudiera haberse completado de otra manera. Los participantes y los líderes locales naturalmente hubieran impulsado la mejor opción de desarrollo a su alcance, incluso si fuera algo que ellos pensaban llevar a cabo de todos modos con los recursos disponibles. Entonces, existe algún otro gasto (inframarginal) que está siendo financiado por la ayuda económica. De manera similar, no se puede descartar la posibilidad de que existan poblaciones fuera del proyecto que se hayan beneficiado de la reasignación del gasto público por parte de las autoridades locales, lo que reduce el impacto medido de participación en el programa.

Este problema es estudiado por van de Walle y Cratty (2005) en el contexto de un proyecto de rutas en zonas rurales de Vietnam. Los autores no encontraron un impacto significativo en las rutas rehabilitadas por el proyecto (con ayuda financiera) en comparación con los proyectos de un grupo de comparación. Esto refleja (en parte) la fungibilidad de la ayuda,

aunque también está presente un sesgo de selección (demostrado por el hecho de que el grado de fungibilidad es exagerado a menos que se controle la asignación geográfica del proyecto).

¿Cómo deben medirse los resultados para las personas pobres? La formulación arquetípica del problema de evaluación descrito en la sección 2 se centra en los impactos medios. Como se mencionó anteriormente, esto incluye el caso en que la medición de resultados toma el valor $Y_i=1$ si la unidad i es pobre y $Y_i=0$ en caso contrario. Esa evaluación generalmente se basa en varias líneas de pobreza, cuyo objetivo es establecer el ingreso mínimo necesario para que la unidad i alcance una utilidad de referencia determinada, que puede interpretarse como el “nivel de vida” mínimo necesario para que un individuo sea considerado “no pobre”. El nivel de utilidad de referencia según la normativa se basa en la capacidad para alcanzar ciertos parámetros, tales como una nutrición, vestimenta y vivienda adecuadas para la actividad física normal y la participación en la sociedad.¹⁶

Con esta reinterpretación de la variable de resultado, el *ATE* y el *TT* proporcionan ahora el impacto del programa según el índice de recuento de la pobreza.(% por debajo de la línea de pobreza). Al repetir los cálculos del impacto para múltiples “líneas de pobreza” se puede hacer un seguimiento del impacto en la distribución acumulativa de los ingresos. También se pueden utilizar las mediciones de la pobreza de orden superior (que penalizan la desigualdad entre los pobres), siempre y cuando formen parte de una clase (amplia) de mediciones aditivas, mediante las cuales la medición de pobreza total pueda expresarse como el promedio poblacional ponderado de todas las mediciones de pobreza individuales en esa población.¹⁷

¹⁶ Se debe tener en cuenta que las líneas de pobreza (en general) varían según los lugares y el tamaño y composición demográfica del hogar, y posiblemente otros factores. Para obtener material sobre teoría y métodos para establecer líneas de pobreza, consulte Ravallion (2006).

¹⁷ Consulte Atkinson (1987) para obtener información sobre la forma general de estas mediciones y ejemplos en la literatura.

Sin embargo, centrarse en los impactos sobre la pobreza no implica que se deba utilizar la variable binaria construida como la variable dependiente (en ecuaciones de regresión tales como (4) ó (5) o especificaciones no lineales como el modelo Probit). Eso significa una pérdida innecesaria de información relevante para explicar por qué algunas personas son pobres y otras no. En lugar de transformar el indicador continuo de asistencia social (según los ingresos o gastos normalizados por la línea de pobreza) en una variable binaria desde el comienzo, probablemente sea mejor aprovechar toda la información disponible en la variable continua, estableciendo repercusiones para la pobreza después del análisis general.¹⁸

¿Qué datos se necesitan? Como queda claro a partir del análisis anterior, la preocupación sobre la utilización de datos inadecuados o inexactos es el quid del problema de evaluación. Al embarcarse en cualquier evaluación de impacto, es importante conocer primero los detalles administrativos/institucionales más importantes del programa; esa información generalmente es proporcionada por la administración del programa. En las evaluaciones NX, dicha información resulta fundamental para diseñar una encuesta que recopile los datos correctos a fin de controlar el proceso de selección. Conocer el contexto y las características de diseño del programa también puede ayudar a manejar la selección de características no observables, ya que en ocasiones esto puede generar restricciones de identificación factibles, tal como se analiza más detenidamente en las secciones 6 y 8.

Las evaluaciones NX pueden resultar difíciles en términos de recopilación de datos y de aplicación de metodología. Existe la tentación de recurrir a entrevistas menos formales y estructuradas con los participantes. No obstante, es difícil formular preguntas contrafactuales en

¹⁸ He escuchado en varias ocasiones que la transformación de la medición del resultado en una variable binaria y la utilización de un modelo Logit o Probit permite obtener un modelo diferente para determinar el nivel de vida de los pobres en relación a los no pobres. Esta afirmación no es correcta, ya que el modelo subyacente en términos de la variable continua latente es el mismo. Los modelos Logit y Probit son sólo estimadores adecuados para ese modelo si la variable continua no se observa, que no es el caso aquí. Para leer más sobre este debate, consulte Ravallion (1996).

entrevistas o en grupos de enfoque. Intente preguntarle a alguien que participa en el programa: “¿qué estaría haciendo ahora si este programa no existiera?” Las conversaciones con los participantes (y los no participantes) pueden ser un complemento valioso para los datos cuantitativos de los estudios, pero es improbable que proporcionen una evaluación de impacto creíble por sí solas.

Los datos sobre resultados y otros determinantes, incluida la participación en el programa, generalmente se obtienen a partir de encuestas. La unidad de observación puede ser una persona, hogar, área geográfica o instalación (escuela o centro para atención de la salud) según el tipo de programa. Los datos de la encuesta se complementan generalmente con otros datos útiles sobre el programa (tales como la base de datos de monitoreo del proyecto).¹⁹

Un problema serio es la comparabilidad de las fuentes de datos sobre los participantes y los no participantes. Las diferencias en el diseño de los instrumentos de encuesta pueden suponer diferencias significativas en las mediciones de resultados. Por ejemplo, Heckman et ál. (1999, Sección 5.33) muestra cómo las diferencias en las fuentes de datos y los supuestos utilizados para el procesamiento de datos pueden provocar importantes diferencias en los resultados obtenidos en la evaluación de programas de capacitación de EE.UU. Diaz y Handa (2004) llegaron a una conclusión similar con respecto al programa PROGRESA de México. Descubrieron que las diferencias en los instrumentos de encuesta generan sesgos significativos en el estimador de correspondencia de puntaje de propensión [*propensity-score matching estimator*] (que se analiza con mayor detalle en la sección 5), a pesar de las buenas aproximaciones a los resultados experimentales logrados utilizando el mismo instrumento de encuesta.

¹⁹ Para consultar excelentes estudios sobre cuestiones generales relacionadas con la recopilación y el análisis de datos de encuestas en hogares en países en desarrollo, consulte Deaton (1995, 1997).

Existen dudas respecto a la exactitud con que las encuestas miden los resultados que generalmente se utilizan en los programas de lucha contra la pobreza. Los totales de consumo e ingresos basados en encuestas para muestras representativas a nivel nacional, no coinciden con los totales obtenidos a partir de las cuentas nacionales (NA). Esto es esperable para el GDP, que incluye fuentes de absorción doméstica que no son hogares. Quizás más sorprendentes sean las discrepancias encontradas en los niveles y tasas de crecimiento del consumo privado en los totales de las NA (Ravallion 2003b).²⁰ Aun así, se debe tener en cuenta que (según se registró en la práctica), el consumo privado en las NA incluye componentes mensurables y de rápido crecimiento que generalmente no se incluyen en las encuestas (Deaton, 2005). Sin embargo, dejando de lado las diferencias en lo que se mide, las encuestas se topan con problemas de subregistro (particularmente para los ingresos; el problema parece ser menos serio para los consumos) y de no respuesta selectiva (la probabilidad de que los ricos respondan es menor).²¹

Los problemas de errores de medición en las encuestas se pueden solucionar, hasta cierto grado, con los mismos métodos que se utilizan para resolver el sesgo de selección. Por ejemplo, si el problema de medición afecta los resultados para las unidades de tratamiento y comparación de manera idéntica (y aditiva) y no están correlacionados con las variables de control, no serán un problema para calcular el efecto medio del tratamiento. Esto señala nuevamente la importancia de los controles. Pero incluso cuando existen variables obvias omitidas correlacionadas con el error de medición, es posible obtener estimaciones confiables utilizando los estimadores de doble diferencia que se describen con mayor detalle en la sección 7. Esto aún requiere que el problema de medición pueda tratarse como un componente de error común (aditivo), que afecte los resultados medidos para las unidades de tratamiento y las de comparación de igual manera. Sin embargo, puede tratarse de supuestos demasiado fuertes en el caso de algunas aplicaciones.

4. Experimentos sociales

²⁰ El grado de discrepancia depende fundamentalmente del tipo de encuesta (especialmente, si recopila datos sobre gastos de consumo o ingresos) y la región; consulte Ravallion (2003b).

²¹ Al medir la pobreza, algunos investigadores han reemplazado el medio de la encuesta por el medio de las cuentas nacionales (GDP o consumo per cápita); consulte, por ejemplo, Bhalla (2002) y Sala-i-Martin (2002). Esto supone que la discrepancia es neutral con respecto a la distribución, lo que es improbable en este caso; por ejemplo, las no respuestas selectivas a las encuestas pueden generar errores no neutrales en absoluto (Korinek et ál., 2005).

Un experimento social tiene como objetivo aleatorizar la implementación, de manera tal que todas las unidades (dentro de un conjunto bien definido) tengan la misma posibilidad *ex-ante* de recibir el programa. La aleatorización incondicional de los programas de lucha contra la pobreza es virtualmente inconcebible, ya que las autoridades responsables de las políticas tienden a asignar dichos programas sobre la base de características observadas, tales como hogares de áreas pobres con muchos dependientes. Es más común que la asignación del programa se realice parcialmente al azar, según ciertas variables observables, X . La implicación clave para la evaluación es que todos los atributos (observados y no observados) anteriores a la intervención sean entonces independientes del hecho de que la unidad reciba o no el programa. Por implicación, $B^{TT} = 0$, y por lo tanto la diferencia *ex-post* observada en los resultados medios entre los grupos de tratamiento y de control es atribuible al programa.²² En términos de la formulación paramétrica del problema de evaluación descrito en la sección 2, la aleatorización garantiza que no exista sesgo de selección en la estimación (3.1) y (3.2) o (de manera equivalente) que el término de error en la ecuación (4) sea ortogonal a todos los regresores. Los no participantes son por lo tanto un grupo de control válido para identificar el contrafáctico,²³ y el impacto medio se calcula (de manera no paramétrica) por la diferencia entre el medio de la muestra de los valores observados de Y_i^T y de Y_i^C (incluidos los valores dados de X_i).

Ejemplos: Varias evaluaciones en los EE.UU. han utilizado aleatorización, con frecuencia aplicada a un esquema piloto. Mucho se ha aprendido sobre reforma de políticas de asistencia social a partir de estas pruebas (Moffitt, 2003). En el caso de programas de mercado laboral activos, dos ejemplos son la Ley de Asociación para la Capacitación Laboral [*Job*

²² Sin embargo, la diferencia simple en los resultados medios no es necesariamente el estimador más eficiente, consulte Hirano et ál. (2003).

²³ El término “grupo de control” con frecuencia se limita a experimentos sociales, mientras que el término “grupo de comparación” se utiliza en evaluaciones NX.

Training Partnership Act o *JTPA*] (consulte, por ejemplo, Heckman et ál., 1997b), y el *US National Supported Work Demonstration* (estudiado por Lalonde, 1986, y Dehejia y Wahba, 1999). En lo relativo a programas de subsidios salariales con objetivos predeterminados en los EE.UU., Burtless (1985), Woodbury y Spiegelman (1987) y Dubin y Rivers (1993) han estudiado evaluaciones aleatorizadas.

Otro ejemplo (bastante diferente) es el experimento Avance hacia la oportunidad (*Moving to Opportunity* o *MTO*), en el que se seleccionaron al azar ocupantes de viviendas sociales en zonas pobres del interior de cinco ciudades de EE.UU. y se les entregaron vales para comprar viviendas en otro lugar (Katz et ál., 2001; Moffitt, 2001). La hipótesis era que los atributos del área de residencia son importantes para las perspectivas individuales de escape de la pobreza. La asignación aleatoria de los vales MTO ayuda a encarar algunas inquietudes de larga data relacionadas con pruebas NX realizadas anteriormente para los efectos de los vecindarios (Manski, 1993).²⁴

También se han realizado varios experimentos sociales en países en desarrollo. Un ejemplo muy conocido es el programa *PROGRESA* de México, que proporciona transferencias de dinero a familias pobres seleccionadas con la condición de que los hijos asistan a la escuela y reciban atención de salud y suplementación nutricional. La (considerable) influencia que este programa ha tenido en la comunidad de desarrollo, indudablemente surge en gran parte gracias al esfuerzo sustancial y público que se puso en esta evaluación. Un tercio de las comunidades muestreadas consideradas elegibles para el programa se eligieron al azar para formar un grupo de control que no recibió el programa durante un período inicial en el que los otros dos tercios sí lo recibían. El acceso público a los datos de la evaluación permitió la realización de varios estudios

²⁴ Se debe tener en cuenta que el diseño del experimento MTO no identifica los efectos del vecindario en el origen, dado que los atributos del destino también tienen relevancia en los resultados (Moffitt, 2001).

valiosos, que señalaron ganancias significativas para el sector de salud (Gertler, 2004), educación (Schultz, 2004; Behrman et ál., 2002) y consumo de alimentos (Hoddinott y Skoufias, 2004). Se puede consultar un estudio completo del diseño, implementación y resultados de PROGRESA en Skoufias (2005).

En otro ejemplo para un país en desarrollo, Newman et ál. (2002) aleatorizaron la elegibilidad para un fondo social financiado por el Banco Mundial destinado a una región de Bolivia. Se descubrió que, dentro del período de evaluación, las inversiones en educación solventadas por el fondo tuvieron un impacto significativo en la infraestructura escolar pero no en los resultados educativos.

La aleatorización también fue utilizada por Angrist et ál. (2002) para evaluar un programa colombiano que asignaba vales de escolaridad mediante un sorteo. Tres años más tarde, los ganadores del sorteo tenían una incidencia significativamente menor de repetición escolar y puntajes más altos en pruebas.

Otro ejemplo es el experimento *Proempleo* de Argentina (Galasso et ál., 2004). Se trató de la evaluación aleatoria de un programa piloto de subsidios salariales y capacitación cuyo objetivo era ayudar a los participantes de programas para desempleados en Argentina a encontrar empleos regulares en el sector privado. Dieciocho meses después, las personas que habían recibido el vale para un subsidio salarial tenían mayores probabilidades de encontrar empleo que el grupo de control. (Volveremos a este tema más adelante en el capítulo, para examinar más detenidamente algunas enseñanzas de esta evaluación).

Algunos sostienen que las agencias de desarrollo como el Banco Mundial deberían hacer un uso más extensivo de estos experimentos sociales. Si bien el Banco Mundial ha apoyado varios experimentos sociales (incluida la mayoría de los ejemplos para países en desarrollo anteriormente mencionados), no ha sido ésta la postura del Departamento de evaluación de

operaciones [*Operation Evaluation Department* u *OED*] del banco (unidad semidependiente para la evaluación *ex-post* de sus propias operaciones de préstamo). De 78 evaluaciones del OED analizadas por Kapoor (2002), sólo una utilizaba aleatorización;²⁵ de hecho, sólo 21 de las evaluaciones utilizaban alguna forma de análisis contrafactual. Cook (2001) y Duflo y Kremer (2005) han recomendado que el OED realice más experimentos sociales.²⁶ Sin embargo, antes de aceptar ese consejo es necesario estar al tanto de algunos de los problemas que presentan los experimentos sociales, los cuales trataremos a continuación.

Problemas relacionados con los experimentos sociales: Se ha debatido mucho si los diseños aleatorios son el modelo ideal para evaluar los programas de lucha contra la pobreza.²⁷ Con frecuencia los experimentos sociales han provocado objeciones éticas y generado sensibilidades políticas, lo que ha retrasado los intentos de implementación, especialmente para los programas del gobierno. Existe la percepción de que los experimentos sociales tratan a las personas como “conejiillos de indias”, negándoles deliberadamente el acceso al programa a algunos que realmente lo necesitan (para formar el grupo de control) y favorecen a otros que no lo necesitan (ya que la asignación aleatoria indudablemente escoge algunas personas que normalmente no participarían del proyecto). En el caso de los programas de lucha contra la pobreza, se termina evaluando impactos para tipos de personas a quienes el programa no estaba dirigido y/o negando el acceso a personas pobres que sí necesitan el programa; en ambos casos en clara oposición al objetivo de combatir la pobreza.

²⁵ Según la descripción de Kapoor no queda claro si esta única evaluación era verdaderamente un experimento social.

²⁶ El OED evalúa sólo proyectos del Banco (incluidas las evaluaciones hechas por el personal del Banco asignado al proyecto) una vez que éstos han finalizado, lo cual dificulta la realización de evaluaciones de impacto adecuadas. Obsérvese que otras unidades del Banco que realizan evaluaciones además del OED, incluido el departamento de investigaciones, utilizan siempre análisis contrafáctico y en ocasiones aleatorización.

²⁷ Para obtener información sobre los argumentos a favor y en contra de los experimentos sociales, consulte (entre otros) Heckman y Smith (1995), Burtless (1995) y Moffitt (2003).

Como se mencionó en la sección 3, la evaluación en sí misma rara vez es la causa de una cobertura incompleta de los pobres dentro de un programa de lucha contra la pobreza, sino que se trata más bien de una cuestión de recursos insuficientes. Cuando existen personas pobres que no pueden ingresar en el programa con los recursos disponibles, se ha argumentado que las objeciones éticas en realidad favorecen los experimentos sociales. De hecho se ha argumentado que la solución más justa para tal situación es asignar el programa de modo aleatorio, para que todas las personas tengan la misma oportunidad de recibir los recursos limitados disponibles.²⁸

El contra argumento es que es difícil apreciar la “justicia” de un programa de lucha contra la pobreza que ignora la información disponible sobre las diferencias en el grado de privación. Una cuestión fundamental aquí es determinar qué es “información disponible”. Por lo general, los experimentos sociales asignan la participación según determinadas características observables. Pero las características que son observables para el evaluador generalmente son un subconjunto de las que están disponibles para los grupos de interés clave. Las preocupaciones éticas con respecto a los experimentos sociales persisten cuando al menos algunos observadores saben que el programa no se ofrece a personas que lo necesitan mientras que sí se ofrece a otras que no lo necesitan.

Se han presentado otras inquietudes relacionadas con los experimentos sociales.

La validez interna puede ser cuestionada cuando existe un cumplimiento selectivo en la asignación aleatoria teórica. Las personas son (generalmente) agentes libres. No están obligados a cumplir con la asignación del evaluador. El hecho de que las personas puedan optar por no participar en la asignación aleatorizada tiende a aliviar el problema ético antes mencionado sobre los experimentos sociales. Las personas que saben que no necesitan el programa supuestamente

²⁸ Según la descripción del estudio de Newman et ál. (2003) parece que éste fue el modo en que se defendió la aleatorización ante las autoridades pertinentes en su caso.

se negarán a participar. Pero el cumplimiento selectivo invalida claramente las inferencias sobre impacto. La magnitud de este problema depende, por supuesto, del programa específico; el cumplimiento selectivo es más probable para (por ejemplo) un programa de capacitación que para un programa de transferencia de dinero. Las secciones 7 y 8 vuelven sobre este tema y analizan cómo los métodos NX pueden ayudar a tratar el problema, y cómo los diseños parcialmente aleatorios pueden ser útiles para identificar impactos con métodos NX.

Los efectos derivados son una causa importante de preocupación acerca de la validez interna de las evaluaciones en la práctica, incluidos los experimentos sociales. Es ampliamente reconocido en la literatura que la selección de las unidades observacionales debe reflejar los posibles efectos derivados. Por ejemplo, Miguel y Kremer (2004) estudiaron la evaluación de tratamientos para gusanos intestinales en niños y argumentan que un diseño aleatorio en el cual algunos niños reciben tratamiento y otros son retenidos para formar un grupo de control, subestimaría seriamente las ganancias del tratamiento al ignorar la externalidades entre los niños “tratados” y los niños del grupo de “control”. El diseño aleatorio del experimento de los autores evitaba este problema utilizando un tratamiento masivo para todas las escuelas en lugar de un tratamiento individual (utilizando escuelas de control a una distancia suficiente de las escuelas bajo tratamiento).

Las respuestas de comportamiento de terceros también pueden generar efectos derivados. Recordemos el ejemplo de la sección 3 sobre cómo la cúpula gubernamental puede ajustar su propio gasto, contrarrestando la asignación (aleatoria o no). Esto puede ser un gran problema para las evaluaciones aleatorias. Es posible que la cúpula gubernamental no considere necesario compensar a las unidades que no ingresaron al programa cuando esto se basó en factores verdaderos y observables reconocidos como relevantes. Por otro lado, las autoridades pueden sentirse obligadas a compensar la “mala suerte” de las unidades asignadas aleatoriamente al

grupo de control. Con la aleatorización pueden darse efectos derivados que no se dan con la selección basada en características observables.

Éste es un ejemplo de un problema más general y fundamental que tienen los diseños aleatorios utilizados en los programas de lucha contra la pobreza, a saber: que el mismo proceso de aleatorización puede alterar la manera en que el programa funciona en la práctica. Es posible que existan diferencias sistemáticas entre las características de las personas normalmente atraídas al programa y aquellas personas de la misma población asignadas aleatoriamente al programa. (Esto es denominado a veces “sesgo de aleatorización”). Heckman y Smith (1995) analizan un ejemplo de la evaluación realizada al programa JTPA, en donde se necesitaron cambios sustanciales en los procedimientos de reclutamiento del programa para formar el grupo de control. El programa piloto evaluado no es entonces el mismo programa que luego se implementa; esto genera dudas sobre la validez de las inferencias que surgen de la evaluación.

El JTPA ilustra otro problema potencial; concretamente, que los factores políticos e institucionales pueden retrasar la asignación aleatoria. Esto es causa de que personas opten por no participar con el consiguiente aumento de costos, ya que se gasta más en candidatos que terminan en el grupo de control (Heckman y Smith, 1995).

Otra crítica a los experimentos sociales señala que, aun con asignación aleatoria, sólo conocemos los resultados medios del contrafáctico, por lo que no se puede inferir la distribución conjunta de los resultados, lo que nos permitiría decir algo (por ejemplo) sobre la proporción de personas que experimentaron ganancias en relación a las que experimentaron pérdidas entre los participantes del programa (Heckman y Smith, 1995). La sección 9 vuelve sobre este tema.

El punto fuerte de los experimentos es que tratan el problema de la implementación dirigida basándose en factores no observables; su punto débil es que no arrojan luz sobre los

determinantes de impactos y otros parámetros relacionados con las políticas, aunque esta debilidad es compartida por muchos métodos NX actualmente en práctica.

¿Qué puede hacerse cuando el programa no fue implementado de manera aleatoria? El resto del capítulo ofrece un estudio crítico de los principales métodos NX que se usan en la práctica.

5. Métodos basados en el puntaje de propensión (*propensity-score*)

Como se hizo hincapié en la sección 3, se prevé que exista un sesgo de selección al comparar una muestra aleatoria de la población de participantes con una muestra aleatoria de no participantes. Existe una presunción generalizada de que tales comparaciones ofrecen información inexacta para la formulación de políticas. Hasta qué grado esto es así, es una cuestión empírica. *A priori*, es preocupante que muchas evaluaciones NX actualmente en práctica proporcionen tan poca información para evaluar si los no participantes del “grupo de comparación” guardan similitud con los participantes en ausencia de la intervención.

Algunos de los sesgos en comparaciones de diferencia simple pueden eliminarse pareando los dos grupos según las características observables. Al tratar de encontrar un grupo de comparación para evaluar el contrafáctico, es natural que se busque a no participantes con características similares a las de los participantes antes de la intervención. Sin embargo, existen muchas características que podrían utilizarse para realizar la comparación. ¿Cómo se deben ponderar las características para seleccionar el grupo de comparación? Esta sección comienza con una revisión teórica y práctica de la comparación utilizando puntajes de propensión. Hacia el final de la sección, también se analizan otros usos de los puntajes de propensión (diferentes a los de comparación) en las evaluaciones.

Correspondencia de puntaje de propensión (Propensity-score matching o PSM): Este método tiene como objetivo seleccionar comparadores basándose en los puntajes de propensión, según resulta de $P(Z) = \Pr(T = 1|Z)$ ($0 < P(Z) < 1$), donde Z es un vector de variables de control anterior a la exposición (que puede incluir valores del indicador de resultados anteriores al tratamiento).²⁹ (Se presume que los valores de Z_i no se ven afectados por el hecho de que la unidad i reciba o no el programa). PSM utiliza $P(Z)$ (o una función monótona de $P(Z)$) para seleccionar unidades de comparación. Rosenbaum y Rubin (1983) demuestran que si los resultados son independientes de la participación dado Z_i , los resultados también son independientes de la participación dado $P(Z_i)$.³⁰ (Esta es una versión más fuerte del supuesto de exogeneidad de implementación analizada en las secciones 2 y 3). La independencia de la condición implica que $B^{TT}(X) = 0$, de manera tal que $E(Y^C|X, T = 1)$ (no observada) pueda reemplazarse simplemente por $E(Y^C|X, T = 0)$ (observada). Por lo tanto, como sucede en un experimento social, el TT no se identifica paramétricamente por la diferencia en los resultados medios de la muestra entre las unidades tratadas y el grupo de comparación pareado ($D(X)$). Según el supuesto de independencia, la correspondencia exacta de $P(Z)$ elimina el sesgo de selección, aunque no proporciona necesariamente el estimador de impacto más eficaz (Hahn, 1998; Angrist y Hahn, 2004).

De manera intuitiva, lo que el PSM hace es crear el análogo observacional de un experimento social en el cual todos tienen la misma probabilidad de participación. La diferencia es que en el PSM la condicional de probabilidad (según Z) es uniforme entre los participantes y

²⁹ El análisis actual se limita al caso estándar de tratamiento binario. Cuando se generaliza para casos de valores múltiples o tratamientos continuos, se define el puntaje de propensión generalizado dado por la probabilidad condicional de un nivel de tratamiento específico (Imbens, 2000; consulte también Hirano e Imbens, 2004).

³⁰ El resultado también requiere que las T_i sean independientes de todas las i . Para una exposición y prueba claras del teorema de Rosenbaum-Rubin consulte Imbens (2004).

los comparadores pareados, mientras que la aleatorización garantiza que los grupos de participantes y de comparación sean idénticos en términos de distribución de todas las características, ya sean observadas o no. El PSM descarta en su supuesto el problema de implementación endógena, dejando sólo la necesidad de equilibrar la probabilidad condicional; es decir, el puntaje de propensión. Una consecuencia de esta distinción es que (a diferencia de un experimento social) las estimaciones de impacto obtenidas por el PSM dependen siempre de las variables utilizadas para la correspondencia y (por lo tanto) de la cantidad y calidad de datos disponibles.

Las variables de control en Z pueden diferir de las covariantes de resultados (el vector X en la sección 2); esta distinción juega un papel importante en la estimación de impacto que se analiza en la sección 8. ¿Pero qué debe incluirse en Z_i ? La teoría de PSM no aporta demasiados datos para responder a esa pregunta, aun así, la elección tiene cierto peso sobre los resultados obtenidos. La elección de variables debe basarse en teoría y/o hechos sobre el programa y su entorno, según su relevancia para la comprensión de los factores económicos, sociales o políticos que afectan la asignación del programa. El trabajo de campo cualitativo puede resultar útil; por ejemplo, las opciones de especificación en Jalan y Ravallion (2003b) reflejaban entrevistas cualitativas con participantes del programa *Trabajar* de Argentina (una combinación de asistencia para desempleados y fondo social) y con los administradores del programa local (con preguntas sobre la participación de las personas en el programa). De manera similar, Godtland et ál. (2004) validaron su elección de covariantes para la participación en un programa de extensión agrícola en Perú mediante entrevistas con agricultores. Claramente, si los datos disponibles no incluyen determinantes importantes de participación, entonces la presencia de características no observables significará que el PSM no es capaz de reproducir (hasta una aproximación razonable) los resultados de un experimento social.

En la práctica se suelen utilizar los valores previstos de una regresión Logit o Probit estándar para estimar el puntaje de propensión de cada observación en las muestras de participantes y no participantes (aunque también se pueden usar modelos de respuesta binaria no paramétricos; consulte Heckman et ál., 1997). La regresión de participación es interesante en sí misma porque puede proporcionar información útil sobre los resultados obtenidos de la asignación de los programas de lucha contra la pobreza (consulte, por ejemplo, el debate en Jalan y Ravallion, 2003b). El grupo de comparación se forma entonces escogiendo el “vecino más cercano” de cada participante, definido como el no participante que minimiza $|\hat{P}(Z_i) - \hat{P}(Z_j)|$ siempre que no exceda un límite ajustable razonable. Teniendo en cuenta los errores de medición, se obtendrán estimaciones más confiables si se toma el resultado medio de (por ejemplo) los cinco vecinos más cercanos, aunque esto no reduce necesariamente el sesgo.³¹ Es una buena idea comprobar la existencia de diferencias sistemáticas en las covariantes entre los grupos de tratamiento y de comparación generadas por el PSM. Smith y Todd (2005a) describen una “prueba de equilibrio” (*balancing test*) muy útil para tal fin.

El estimador del PSM típico para el impacto medio se expresa como

$\sum_{j=1}^{NT} (Y_j^T - \sum_{i=1}^{NC} W_{ij} Y_{ij}^C) / NT$ en donde NT es la cantidad de personas que reciben el programa, NC la cantidad de no participantes y W_{ij} las ponderaciones. Se han utilizado varios esquemas de ponderación, desde el sistema del vecino más próximo hasta ponderaciones no paramétricas basadas en funciones *kernel* de las diferencias en los puntajes, según los cuales todas las unidades de comparación se utilizan para generar el contrafáctico para cada unidad que participa en el programa, pero con una ponderación que alcanza su valor máximo para el vecino más

³¹ Rubin y Thomas (2000) utilizan simulaciones para comparar el sesgo al utilizar cinco vecinos próximos y al utilizar sólo uno. No se registraron patrones definidos.

próximo y se reduce a medida que aumenta la diferencia absoluta en puntaje de propensión; Heckman et ál. (1997b) analiza este esquema de ponderación.³²

Las propiedades estadísticas de los estimadores de correspondencia (en particular sus propiedades asintóticas) aún no se comprenden plenamente. En la práctica, los errores estándar surgen normalmente mediante un método de muestreo repetitivo (*bootstrapping*), aunque no es evidente que este método sea el apropiado en todos los casos. Abadie e Imbens (2006) examinan las propiedades formales en muestras grandes de estimadores de correspondencia de vecino K más cercano (para los cuales el método estándar de *bootstrapping* no produce errores estándar válidos) y proporciona un estimador coherente para el error asintótico estándar.

Los impactos medios también se pueden calcular según las características observadas. En los programas de lucha contra la pobreza, se busca comparar el impacto medio condicional entre diferentes ingresos anteriores a la intervención. Para cada participante que forma parte de una muestra, se estima el incremento en los ingresos a partir del programa comparando los ingresos de ese participante con los ingresos de un no participante pareado. Al restar la ganancia estimada de los ingresos observados luego de la intervención, se puede estimar qué lugar ocuparía cada participante en la distribución de ingresos sin el programa. Si se promedia este valor entre diferentes estratos definidos por los ingresos anteriores a la intervención se puede evaluar la incidencia de los impactos. Es una buena idea comprobar si los puntajes de propensión (e incluso las mismas Z) están correctamente equilibradas dentro de los estratos (y en el total), ya que se corre el riesgo de confundir errores de comparación con efectos reales.

³² Frölich (2004) compara propiedades finitas de muestra de varios estimadores y descubre que el método de regresión local lineal contraída es más eficiente y sólido que otros métodos alternativos.

De manera similar, se pueden construir las funciones de distribución acumulativa empíricas y contrafactuales, o las integrales empíricas, y comprobar el predominio sobre un rango relevante de líneas y mediciones de pobreza. Esto se ilustra en la Figura 2 para el programa *Trabajar* de Argentina. La figura proporciona la función de distribución acumulativa (CDF por sus siglas en inglés) (o “curva de incidencia de la pobreza”) mostrando cómo el índice de recuento de la pobreza (% por debajo de la línea de pobreza) varía dentro de un amplio rango de posibles líneas de la pobreza (cuando ese rango cubre todos los ingresos, se obtiene la función de distribución acumulativa estándar). La línea vertical es una línea de pobreza ampliamente utilizada para Argentina. La figura muestra además el CDF contrafactual estimado, después de restar los incrementos de ingresos imputados a los ingresos observados (posteriores a la intervención) en todos los participantes que formaban parte de la muestra. Utilizando una línea de pobreza de \$100 por mes (según la cual cerca del 20% de la población nacional es considerada pobre) se observó una reducción de 15 puntos porcentuales en la incidencia de la pobreza entre los participantes gracias al programa; esto aumenta a 30 puntos porcentuales si se utilizan líneas de pobreza más cercanas al estrato inferior de la distribución. También se puede observar el aumento en cada percentil de la distribución (mirando horizontalmente) o el impacto en la incidencia de la pobreza según cualquier línea de pobreza (mirando verticalmente).³³

Al evaluar programas de lucha contra la pobreza en países en desarrollo, las comparaciones de diferencia simple al utilizar PSM ofrecen la ventaja de no requerir aleatorización ni datos iniciales (previos a la intervención). Si bien esto puede ser una gran ventaja en la práctica, también tiene su costo. Para aceptar el supuesto de exogeneidad se debe asegurar el control de los factores que influyen de manera conjunta la implementación y los

³³ En Ravallion (2003b) se pueden consultar otros debates sobre cómo los resultados de una evaluación de impacto por PSM se pueden utilizar para evaluar impactos en las mediciones de pobreza **en lugar de otras mediciones y la línea de pobreza**.

resultados del programa. En la práctica, se debe considerar la posibilidad de que exista una variable latente que influya de manera conjunta la implementación y los resultados (invalidando, por lo tanto, el supuesto clave de independencia condicional del PSM). Esto debe decidirse para la aplicación con la que se está trabajando en ese momento. La sección 7 ofrecerá ejemplos de hasta dónde puede llegar el método con datos inadecuados sobre covariantes combinadas de participación y resultados.

¿En qué se diferencia el PSM de otros métodos? En un experimento social (al menos en su forma pura), el puntaje de propensión es una constante ya que todos tienen la posibilidad de recibir el tratamiento. La asignación aleatoria asegura que las distribuciones de observables y no observables se equilibren entre las unidades de tratamiento y las de comparación. Por el contrario, el PSM sólo intenta equilibrar la distribución de observables; de allí las inquietudes sobre el sesgo de selección en las estimaciones de este método. Tampoco se puede asumir que la eliminación del sesgo de selección basado en observables reducirá el sesgo total; eso sólo sucedería si dos fuentes de sesgo –que asociadas con observables y causadas por factores no observados– fueran en la misma dirección, lo cual no se puede garantizar *a priori*. Si el sesgo de selección basado en no observables contrarresta al basado en observables, entonces eliminar sólo el último sesgo aumentará el sesgo total. Si bien esto es posible en teoría, los estudios de replicación (que comparan evaluaciones NX con experimentos para los mismos programas) no parecen haber encontrado un ejemplo en la práctica. Más adelante analizo lecciones sobre estudios de replicación.

Lo habitual es comparar el PSM y la regresión OLS de los indicadores de resultado sobre variables ficticias (*dummy*) para la implementación del programa, lo que permite introducir las covariantes observables como controles lineales (como en el caso de las ecuaciones 4 y 5). OLS requiere esencialmente el mismo supuesto de independencia condicional

(exogeneidad) que el PSM, pero también impone supuestos de forma funcional arbitrarios relacionados con los efectos del tratamiento y las variables de control. Por otro lado, el PSM (al igual que los métodos experimentales) no requiere un modelo paramétrico que vincule los resultados con la participación en el programa. Por lo tanto, el PSM permite estimar los impactos medios sin supuestos arbitrarios sobre formas funcionales y errores de distribución. Esto también puede facilitar la comprobación de posibles efectos de interacción compleja. Por ejemplo, Jalan y Ravallion (2003a) utilizan el PSM para estudiar cómo los efectos de interacción entre ingresos y educación afectan los beneficios de salud de los niños a partir del acceso a agua potable en zonas rurales de la India. Los autores encontraron un complejo patrón de efectos de interacción. Por ejemplo, la pobreza atenúa los beneficios de salud del agua potable, pero en menor medida cuando mayor es el nivel de educación maternal.

El PSM también difiere de los métodos de regresión estándar en lo relativo a la muestra. Con el PSM la atención se centra en la región de soporte común (Figura 1). Los no participantes con un puntaje menor al de los participantes quedan excluidos. También se pueden restringir las posibles correspondencias de otras maneras, según el entorno. Por ejemplo, se pueden restringir las correspondencias para que comiencen dentro de la misma área geográfica, a fin de asegurar que las unidades de comparación provengan del mismo entorno económico. Por el contrario, los métodos de regresión que se mencionan comúnmente en la literatura utilizan la totalidad de la muestra. Las simulaciones en Rubin y Thomas (2000) indican que las estimaciones de impacto basadas en muestras completas (sin parear) por lo general tienen mayor sesgo, y son más sensibles a las especificaciones incorrectas de la función de regresión que las basadas en muestras pareadas.

Otra diferencia está relacionada con la opción de las variables de control. En el método de regresión estándar, se buscan elementos que permitan predecir los resultados y se da

preferencia a las variables que puedan considerarse exógenas a los resultados. En el PSM, en cambio, se buscan covariantes de participación, que posiblemente incluyan variables que no son buenos indicadores de predicción de resultados. De hecho, resultados analíticos y simulaciones indican que incluso las variables con poca capacidad predictiva de resultados pueden ayudar a reducir el sesgo al estimar los efectos causales usando el PSM (Rubin y Thomas, 2000).

Cuál sería la diferencia en el impacto medio al utilizar el PSM en lugar de OLS es una cuestión empírica. Se han realizado muy pocos estudios metodológicos comparativos. Una excepción es el de Godtland et ál. (2004), donde se utiliza un resultado de regresión y un PSM para evaluar el impacto de las escuelas rurales en el conocimiento de los campesinos sobre las mejores prácticas para el manejo de plagas en el cultivo de papas. Los investigadores informaron que sus resultados fueron convincentes para el cambio del método utilizado.

¿Es bueno el rendimiento del PSM? Utilizando el mismo conjunto de datos del estudio de Lalonde (1986) (descrito en la sección 3), Dehejia y Wahba (1999) descubrieron que el PSM lograba una aproximación bastante buena, mucho mejor que la de los métodos NX estudiados por Lalonde. Al parecer, el rendimiento poco satisfactorio de los métodos NX utilizados por Lalonde se debe en gran parte al uso de unidades observacionales fuera de la región de soporte común. No obstante, la solidez de los hallazgos de Dehejia y Wahba relacionados con la selección de muestras y la especificación elegida para calcular los puntajes de propensión fue cuestionada por Smith y Todd (2005a), quienes sostienen que el PSM no resuelve el problema de selección en el programa estudiado por Lalonde.³⁴

Intentos similares de comprobación del PSM, comparándolo con evaluaciones aleatorias, han arrojado resultados variados. Agodini y Dynarski (2004) no encuentran evidencia

³⁴ Dehejia (2005) responde a Smith y Todd (2005a), quien ofrece una replica en Smith y Todd (2005b). Consulte también Smith y Todd (2001).

consistente de que el PSM pueda replicar los resultados experimentales a partir de sus evaluaciones a programas de abandono escolar en los EE.UU. Utilizando la base de datos de *PROGRESA*, Diaz y Handa (2004) sostienen que el PSM tiene un buen rendimiento, siempre y cuando se utilice el mismo instrumento de encuesta para medir resultados de los grupos de tratamiento y de comparación. Heckman et ál. (1997a, 1998) también enfatizan la importancia de usar el mismo instrumento de encuesta en PSM dentro del contexto de su evaluación a un programa de capacitación de EE.UU. Este último estudio señala además la importancia de que los participantes y no participantes provengan de los mismos mercados laborales locales, y la posibilidad de controlar los antecedentes laborales. El metaestudio de Glazerman et ál. (2003) sostiene que el PSM es uno de los métodos NX capaces de reducir significativamente el sesgo, particularmente cuando se utiliza en combinación con otros métodos.

Otros usos del puntaje de propensión en las evaluaciones: Existen otros métodos de evaluación que utilizan el puntaje de propensión. Estos métodos pueden resultar ventajosos en comparación con el PSM, aunque se han aplicado en muy pocas oportunidades a programas de lucha contra la pobreza en países en desarrollo.

Si bien la correspondencia basada en puntajes de propensión elimina el sesgo (según el supuesto de exogeneidad condicional), no es necesariamente el método de estimación más eficiente (Hahn, 1998). Hirano et ál. (2003) proponen un estimador de impacto alternativo a la correspondencia de puntajes de propensión. Este método pondera las unidades de observación por los inversos de una estimación no paramétrica de los puntajes de propensión. Hirano et ál. demuestran que esta práctica produce un estimador totalmente eficaz para los efectos medios del tratamiento. Chen et ál. (2006) proporcionan una aplicación en el contexto de una evaluación de impacto a largo plazo sobre la pobreza en un programa de desarrollo para áreas pobres de China.

Los puntajes de propensión también pueden utilizarse en el contexto de estimadores basados en una regresión más estándar. Supongamos que se agregó simplemente el puntaje de propensión estimado $\hat{P}(Z)$ a una regresión OLS de la variable de resultado en la variable ficticia de tratamiento, T . (También se puede incluir un efecto de interacción entre $\hat{P}(Z_i)$ y T_i .) Según los supuestos de PSM, esto eliminará cualquier sesgo variable omitido al haber excluido Z de esa regresión, siendo que Z es independiente del tratamiento proporcionado $P(Z)$.³⁵ Sin embargo, este método no tiene la flexibilidad no paramétrica de PSM. Agregar una función adecuada de $\hat{P}(Z)$ al resultado de regresión es un ejemplo del enfoque de “control de función” (CF por su sigla en inglés), en donde bajo condiciones estándar (lo que incluye la exogeneidad de X y Z), el término de sesgo de selección puede expresarse como una función de $\hat{P}(Z)$.³⁶ La identificación se apoya en la no linealidad de CF en Z o en la existencia de una o más covariantes de participación (el vector Z) que sólo afecta los resultados *mediante* la participación. Sujeta esencialmente a las mismas condiciones de identificación, otra opción es utilizar $\hat{P}(Z)$ como variable instrumental para la implementación del programa; opción que también se analiza más detalladamente en la sección 8.

6. Cómo aprovechar el diseño del programa

En ocasiones, los estimadores NX pueden aprovechar al máximo las características del diseño del programa para la identificación. Las discontinuidades generadas por el criterio de elegibilidad del programa pueden ayudar a identificar impactos en las proximidades

³⁵ Esto proporciona una profundización sobre la manera en que funciona PSM; consulte el debate en Imbens (2004).

³⁶ Heckman y Robb (1985) proporcionan un análisis exhaustivo de este enfoque; consulte además el análisis de Heckman y Hotz (1989). Para obtener información sobre la relación entre CF y PSM, consulte Heckman y Navarro-Lozano (2004) y Todd (2006). Sobre la relación entre los enfoques que utilizan CF y los estimadores de variables instrumentales (analizadas con mayor detalle en la sección 8); consulte Vella y Verbeek (1999).

de los puntos de corte para elegibilidad. Las demoras en la implementación de un programa también pueden facilitar la formación de grupos de comparación, lo que puede ayudar a captar algunas fuentes de heterogeneidad latente.

Diseños de discontinuidad: Bajo ciertas condiciones, es posible inferir impactos a partir de las diferencias en los resultados medios entre unidades de uno u otro lado del punto de corte que determina la elegibilidad del programa. Para ver con mayor claridad el funcionamiento de este método, digamos que M_i expresa el puntaje recibido por la unidad i en una prueba de calificación de socioeconómica (*proxy-means test*) y que m expresa el punto de corte de elegibilidad, de manera tal que $T_i = 1$ para $M_i \leq m$, y $T_i = 0$ en caso contrario. Los ejemplos incluyen una prueba de calificación socioeconómica que establece un puntaje máximo de elegibilidad (sección 3) y programas que limitan la elegibilidad a límites geográficos determinados. El estimador de impacto es $E(Y^T | M = m - \varepsilon) - E(Y^C | M = m + \varepsilon)$ para algunas $\varepsilon > 0$ arbitrariamente pequeñas. En la práctica, existe cierto grado inevitable de discrepancia en la aplicación de las pruebas de elegibilidad. Por lo tanto, en lugar de suponer aplicación y cumplimiento rigurosos, se puede seguir la propuesta de Hahn et ál. (2001) de utilizar una probabilidad de participación en el programa, $P(M) = E(T|M)$, que es una función incremental de M con una discontinuidad en m . La idea básica sigue siendo la misma, es decir, que los impactos se miden por la diferencia entre los resultados medios en las proximidades de m .

El supuesto de identificación clave para este estimador es que no existe discontinuidad en los resultados contrafactuales en m .³⁷ El hecho de que un programa tenga reglas de elegibilidad más o menos estrictas no significa (de por sí) que se trate de un supuesto

³⁷ Hahn et ál. (2001) proporcionan un análisis formal de identificación y estimación de impactos para diseños de discontinuidad basados en este supuesto.

factible. Por ejemplo, los límites geográficos para la elegibilidad del programa con frecuencia coinciden con jurisdicciones políticas locales, lo que implica diferencias geográficas actuales o pasadas en (por ejemplo) políticas e instituciones fiscales locales que pueden entorpecer la identificación. La factibilidad del supuesto de continuidad para los resultados contrafactuales debe determinarse en cada aplicación.

En una prueba destinada a comprobar la capacidad de los diseños de discontinuidad para reducir el sesgo de selección, Buddelmeyer y Skoufias (2004) utilizan los puntos de corte de las reglas de elegibilidad de *PROGRESA* para medir los impactos y comparar los resultados con aquellos obtenidos al aprovechar el diseño aleatorio del programa. Los autores descubrieron que el diseño de discontinuidad proporciona buenas aproximaciones para casi todos los indicadores de resultado.

Sin embargo, este método también tiene sus desventajas. Se da por supuesto que el evaluador conoce M_i y, por lo tanto, la elegibilidad para el programa. Pero éste no será siempre el caso. Considere (nuevamente) una transferencia de ingresos en la que los ingresos de los participantes sean inferiores a un punto de corte predeterminado. En una encuesta transversal, se observan los ingresos de los participantes después del programa y los ingresos de los no participantes; pero, por lo general, no se conocen los ingresos en el momento de realizar la prueba. Si estimáramos la elegibilidad restando el pago de la transferencia de los ingresos observados, estaríamos suponiendo (de manera implícita) exactamente lo que deseamos comprobar: si hubo una respuesta conductual respecto al programa. Resulta útil realizar preguntas retrospectivas sobre los ingresos durante la ejecución de la prueba de ingresos (teniendo en cuenta los posibles sesgos), así como llevar a cabo una encuesta inicial o cercana a la fecha de realización de la prueba. Una encuesta inicial también puede ayudar a eliminar

diferencias previas a la intervención en los resultados de ambos lados de la discontinuidad, en cuyo caso se puede combinar el diseño de discontinuidad con el método de doble diferencia, analizado más detenidamente en la sección 7.

Cabe destacar que un diseño de discontinuidad estima el impacto medio para una muestra seleccionada de participantes, mientras que la mayoría de los otros métodos (como los experimentos sociales y el PSM) proporciona el impacto medio para el grupo de tratamiento en su totalidad. Sin embargo, el problema de soporte común anteriormente mencionado, generado en ocasiones por el criterio de elegibilidad, puede significar que otras evaluaciones también se limitan a una submuestra altamente seleccionada; la pregunta es, entonces, si se trata de una submuestra interesante. El truncamiento de las muestras del grupo de tratamiento garantiza un soporte común que probablemente tenderá a excluir a los individuos con mayor probabilidad de participación (para los cuales es más difícil encontrar comparadores de no participación), mientras que los diseños de discontinuidad tenderán a incluir sólo a aquellos con menor probabilidad de participación. Esta última submuestra puede, no obstante, ser relevante para decidir la expansión del programa; la sección 9 vuelve sobre este punto.

Aunque los impactos en las proximidades del punto de corte se identifican de manera no paramétrica en los diseños de discontinuidad, la literatura aplicada ha usado con mayor frecuencia un método paramétrico alternativo en el cual la discontinuidad en el criterio de elegibilidad se utiliza como una variable instrumental para la implementación del programa. Volveremos a este tema a fin de proporcionar ejemplos en la sección 8.

Comparaciones iniciales [Pipeline Comparisons]: La idea aquí es utilizar como grupo de comparación a las personas que se postularon para el programa pero no lo recibieron.³⁸

³⁸ En ocasiones en la literatura esta práctica se denomina “pareamiento inicial”, aunque este término no es adecuado ya que en realidad no se realiza ningún tipo de pareamiento.

PROGRESA es un ejemplo; una tercera parte de los participantes elegibles no recibieron el programa durante 18 meses, y durante ese lapso de tiempo formaron el grupo de control. En el caso de *PROGRESA*, la comparación inicial se hizo de manera aleatoria. También se han utilizado comparaciones iniciales NX en países en desarrollo. Un ejemplo es el de Chase (2002), que utilizó comunidades que se habían postulado para un fondo social (en Armenia) para organizar el grupo de comparación a fin de estimar los impactos del fondo en las comunidades que efectivamente recibieron la asistencia. En otro ejemplo, Galasso y Ravallion (2004) evaluaron un amplio programa de protección social ofrecido por el gobierno de Argentina, el llamado *Plan Jefes y Jefas*, que fue la principal respuesta en política social a la grave crisis económica sufrida en el año 2002. A fin de formar un grupo de comparación para los participantes, se valieron de personas que se habían postulado con éxito para el programa pero que aún no lo habían recibido. Cabe destacar que este método resuelve hasta cierto punto el problema de heterogeneidad latente presente en otros estimadores de diferencia simple, tales como el PSM; el proceso de selección anterior hará que los postulantes exitosos tengan probablemente características no observadas similares, hayan recibido el tratamiento o no.

El supuesto clave aquí es que el período de tratamiento es aleatorio respecto de la aplicación. En la práctica, se debe anticipar un posible sesgo surgido del tratamiento selectivo de los postulantes o de las respuestas conductuales de los postulantes que esperan recibir el tratamiento. Éste es un problema con mayor relevancia en algunos entornos que en otros. Por ejemplo, Galasso y Ravallion argumentaban que, en su caso, esto no había sido un problema ya que realizaron su evaluación del programa durante un período de rápido incremento progresivo, durante la crisis financiera del año 2002 en Argentina, cuando era materialmente imposible ayudar inmediatamente a todas las personas necesitadas. Los autores también comprobaron las diferencias observables entre dos subconjuntos de postulantes, y encontraron que los observables

(incluido el impacto idiosincrásico en los ingresos durante la crisis) estaban bien equilibrados entre los dos grupos, lo que alivió las inquietudes sobre posibles sesgos. También resultó útil el uso de observaciones longitudinales; volveremos a este ejemplo en la siguiente sección.

Cuando son factibles, las comparaciones iniciales ofrecen un estimador de impacto de diferencia simple que probablemente sea más sólido ante la heterogeneidad latente. Sin embargo se debe comprobar que las estimaciones no contengan sesgos de selección basados en características observables y (si es necesario), se puede utilizar un método como PSM para eliminar la heterogeneidad observable antes de realizar la comparación inicial (Galasso y Ravallion, 2004).

Las comparaciones iniciales también pueden combinarse con diseños de discontinuidad. Aunque no he visto su uso en la práctica, una posible estrategia de identificación para proyectos que se expanden a lo largo de una ruta bien definida es medir los resultados a ambos lados del límite del proyecto actual. Los ejemplos pueden incluir proyectos que conectan de manera progresiva casas con redes de agua potable, cloacas, transporte o comunicaciones existentes, junto con proyectos que expanden esa red en diferentes etapas. Con frecuencia las nuevas instalaciones (electricidad y telecomunicaciones) se expanden junto con las redes de infraestructura preexistentes (tales como calles, para tender el cableado a lo largo de su servidumbre de paso). Por supuesto, también se deben tener en cuenta los efectos de la heterogeneidad observable y del tiempo. También puede existir inquietud sobre los efectos derivados; la conducta de los no participantes puede cambiar ante la posibilidad de ser conectados a la red en expansión.

7. Diferencias de orden superior

Hasta el momento, el análisis se ha centrado en varios estimadores de diferencia simple, que solamente exigen una encuesta transversal adecuada. Podemos obtener más información si realizamos un seguimiento tanto de los participantes como de los no participantes en un período

que se considere suficiente como para captar el impacto de la intervención. La disponibilidad de un punto de partida previo a la intervención, en el que se sabe finalmente quién participa y quién no, puede revelar problemas de especificación en el estimador NX de diferencia simple. Por ejemplo, si la regresión de los resultados (como en las ecuaciones 4 ó 5) se especifica correctamente, la ejecución de la regresión en los datos iniciales debería indicar una estimación de impacto medio que no sea significativamente diferente a cero (Heckman y Hotz, 1989).

Sin embargo, con datos iniciales también se puede estimar el impacto en virtud de un supuesto más débil que la exogeneidad condicional ($B^{TT} = 0$). Esta sección en primer lugar analiza el muy utilizado método de doble diferencia (DD), que se vale de un punto de partida previo a la intervención y al menos una encuesta de seguimiento (posterior a la intervención). El análisis luego se orienta a situaciones –comunes en la evaluación de programas de redes de seguridad que se organizan rápidamente como respuesta a una crisis– en las que no es posible realizar una encuesta inicial, pero se puede llevar a cabo un seguimiento de los ex participantes. Este es un ejemplo de estimador de triple diferencia.

Estimador de doble diferencia: Es un enfoque muy utilizado para abordar las cuestiones relacionadas con la implementación endógena en comparaciones transversales de diferencia simple. La idea básica es comparar muestras de participantes y no participantes antes y después de la intervención. Después de la encuesta inicial de no participantes y de participantes (subsiguientes), se realiza una encuesta de seguimiento de ambos grupos después de la intervención. Finalmente, se calcula la diferencia entre los valores del “antes” y el “después” de los resultados medios correspondientes a cada uno de los grupos de tratamiento y de comparación. La diferencia entre estas dos diferencias medias (de ahí el nombre de “doble diferencia” o “diferencia en la diferencia”) constituye la estimación del impacto.

Para ver de qué se trata en términos más formales, supongamos que Y_{it} expresa la medición de resultados correspondiente a la unidad de observación i observada en dos fechas, $t=0,1$. Por definición, $Y_{it} = Y_{it}^C + T_{it}G_{it}$ y (como en el problema arquetípico de evaluación descrito en la sección 2) se da por supuesto que podemos observar T_{it}, Y_{it}^T cuando $T_{it} = 1$, Y_{it}^C para $T_{it} = 0$, pero que $G_{it} = Y_{it}^T - Y_{it}^C$ no es directamente observable para ninguna i (o en expectativa) porque nos faltan datos de Y_{it}^T para $T_{it} = 0$ y de Y_{it}^C para $T_{it} = 1$. Para resolver el problema de la “falta de datos”, el estimador *DD* supone que el sesgo de selección (la diferencia no observada en los resultados medios contrafactuales entre unidades tratadas y no tratadas) no varía con el tiempo, en cuyo caso los cambios en los resultados correspondientes a los no participantes revelan los cambios en los resultados contrafactuales, es decir que:

$$E(Y_1^C - Y_0^C | T_1 = 1) = E(Y_1^C - Y_0^C | T_1 = 0) \quad (8)$$

Está claro que es un supuesto más débil que el de la exogeneidad condicional en estimaciones de diferencia simple; $B_t^{TT} = 0$ para todos los t implica (8) pero no es necesario para (8). Dado que el período 1 es el inicio, con $T_{0i} = 0$ para todas las i (por definición), $Y_{0i} = Y_{0i}^C$ para todas las i . Queda claro, entonces, que el estimador de doble diferencia arroja el efecto del tratamiento medio sobre los tratados durante el período 1:

$$DD = E(Y_1^T - Y_0^C | T_1 = 1) - E(Y_1^C - Y_0^C | T_1 = 0) = E(G_1 | T_1 = 1) \quad (9)$$

Cabe destacar que los datos del panel no son necesarios para calcular con el método *DD*. Todo lo que se necesita es el conjunto de los cuatro promedios que constituyen la *DD*. Los promedios no necesitan calcularse sobre la misma muestra a lo largo del tiempo.

Cuando los promedios contrafactuales no varían con el tiempo ($E[Y_1^C - Y_0^C | T_1 = 1] = 0$), las ecuaciones (8) y (9) se despliegan en una comparación reflexiva en que solamente se

controlan los resultados correspondientes a las unidades del tratamiento. El hecho de que los resultados medios contrafactuales no cambien es un supuesto poco probable en la mayoría de las aplicaciones. Sin embargo, con suficientes observaciones a lo largo del tiempo, los métodos para probar si hay cambios estructurales en las series de tiempo de los resultados correspondientes a los participantes, pueden ofrecer cierta esperanza de identificar impactos. Para ver un ejemplo, consulte Piehl et ál. (2003).

Para calcular errores estándar e implementar estimadores ponderados (que pueden ayudar a resolver posibles sesgos en la *DD*, tal como se analiza a continuación) es conveniente usar un estimador de regresión para *DD*. Se combinan los datos correspondientes a los dos períodos y al estado del tratamiento y se ejecuta la regresión:

$$Y_{it} = \alpha + DD.T_{it}t + \gamma T_{it} + \delta t + \varepsilon_i \quad (t = 0,1; i = 1, \dots, n) \quad (10)$$

Cabe destacar que es el coeficiente de $T_{it}t$ el que arroja el estimador de impacto medio. Sin embargo, T_{it} debe incluirse como un regresor diferente para captar cualquier diferencia en el promedio de los efectos individuales latentes entre las unidades de tratamiento y de comparación, como surgiría de un sesgo de selección intencional inicial en el programa.³⁹ Cabe destacar (nuevamente) que (10) no requiere datos de panel.

El estimador *DD* puede generalizarse de inmediato a varios períodos y la *DD* puede calcularse, entonces, mediante la regresión de Y_{it} sobre la variable ficticia de participación (individual y específica en cuanto a la fecha) T_{it} , con efectos individuales y fijos en el tiempo.⁴⁰

³⁹ Esto es equivalente a un estimador de efectos fijos en el que el término de error comprende un efecto individual latente que está potencialmente correlacionado con el estado del tratamiento.

⁴⁰ Como es bien sabido, a la hora de calcular los errores estándar del estimador *DD* debe tenerse en cuenta si el término de error diferenciado se correlaciona en serie; Bertrand et ál. (2004) demuestran la posibilidad de que haya grandes sesgos en los errores estándar (OLS) sin corregir de los estimadores *DD*.

Ejemplos de evaluaciones DD: Duflo (2001) calculó el impacto de la construcción de escuelas sobre la educación y los ingresos en Indonesia. Una característica del mecanismo de asignación fue que se construyeron más escuelas en lugares con bajos índices de inscripción. Además, los grupos etarios que participaban en el programa podían identificarse fácilmente. El hecho de que los aumentos en los logros escolares de los primeros grupos expuestos al programa fueran mayores en las áreas que recibieron más escuelas se tomó como indicador de que construir escuelas promovía una mejor educación. Frankenberg et ál. (2005) utilizan un método similar para evaluar los impactos de proporcionar servicios básicos de salud relacionados con el estado nutricional de los niños (la altura según la edad) a través de parteras, también en Indonesia.

En otro ejemplo, Galiani et ál. (2005) usaron un diseño *DD* para calcular el impacto de la privatización de los servicios de agua sobre la mortalidad infantil en Argentina. Para identificar los impactos, los autores aprovecharon la variación intertemporal y geográfica (en distintas municipalidades) en conjunto, tanto en la mortalidad infantil como en la propiedad de los servicios de agua. Los resultados sugieren que la privatización de los servicios de agua redujo la mortalidad infantil.

También se puede usar un diseño *DD* para corregir posibles sesgos en un experimento social, donde hay cierta forma de cumplimiento selectivo u otra distorsión en la asignación aleatoria (tal como se expuso en la sección 4). Se puede encontrar un ejemplo en Thomas et ál. (2003), quienes aleatorizaron la asignación de un suplemento de hierro en píldoras en Indonesia, con un grupo no aleatorizado que recibió un placebo. Al recabar datos iniciales previos a la intervención en ambos grupos, los autores pudieron abordar las preocupaciones respecto al sesgo de cumplimiento.

Si bien el estimador *DD* de diseño clásico realiza un seguimiento de las diferencias en el transcurso del tiempo entre participantes y no participantes, ésta no es la única posibilidad. Jacoby (2002) usó un diseño *DD* para probar si la asignación de recursos dentro del hogar cambiaba en respuesta a un programa de alimentación en la escuela, para neutralizar el efecto de éste en la nutrición infantil. Algunas escuelas tenían el programa de alimentación y otras no, y algunos niños asistían a la escuela mientras que otros no lo hacían. La estimación del impacto *DD* del autor fue la diferencia entre la ingesta calórica media de los niños que habían asistido (el día anterior) a una escuela que tenía programa de alimentación y la media correspondiente a quienes no habían asistido a dichas escuelas, menos la diferencia correspondiente entre los niños que asistieron y los que no asistieron a la escuela hallada en las escuelas que no tenían el programa.

Otro ejemplo se puede encontrar en Pitt y Khandker (1998), quienes evaluaron el impacto de la participación en el Grameen Bank (GB) de Bangladesh sobre varios indicadores relevantes para los estándares de vida actuales y futuros. El crédito del GB está destinado a grupos familiares sin tierra que viven en poblaciones pobres. Algunas de las poblaciones incluidas en el muestreo no reunían los requisitos para el programa y dentro de las poblaciones que sí los reunían, algunos grupos familiares no podían participar debido a que tenían tierra (aunque no está claro hasta qué punto esto se respetó). Los autores tácitamente usaron un diseño *DD* poco frecuente para calcular el impacto.⁴¹ Por supuesto que los retornos por poseer tierra son más altos en poblaciones que no tienen acceso al crédito del GB (dado que el acceso al GB aumenta los retornos de quienes no tienen tierra). Comparar los retornos por poseer tierra entre dos grupos

⁴¹ En mi opinión; Pitt y Khandker (1998) no hacen referencia a la interpretación de su diseño *DD*. Sin embargo, es fácil verificar que el estimador de impacto implícito en la resolución de las ecuaciones (4a-d) de su trabajo es el estimador *DD* que se describe aquí. (Cabe destacar que, para obtener el parámetro del impacto del GB, el *DD* resultante debe normalizarse por medio de la proporción de grupos familiares sin tierra de poblaciones que reúnen los requisitos para participar).

de poblaciones idénticos –salvo por el hecho de que uno reúne los requisitos para el GB y el otro no– revela el impacto del crédito del GB. Por lo tanto, la estimación de Pitt-Khandker del impacto del GB en realidad representa el impacto sobre los retornos a la tierra que surge de retirar el acceso al GB a nivel de la población.⁴² Por inferencia, el “punto de partida previo a la intervención” en el estudio de Pitt-Khandker está dado por las poblaciones que tienen el GB, y el “programa” que se evalúa no es el GB sino más bien el hecho de tener tierra y, por lo tanto, de no reunir los requisitos para el GB. (Volveré a este ejemplo más adelante).

El uso de diferentes métodos y conjuntos de datos en el mismo programa puede resultar revelador. En comparación con el estudio de Jalan y Ravallion (2002b) sobre el mismo programa (programa *Trabajar* de Argentina), Ravallion et ál. (2005) usaron un instrumento de encuesta más ligero, con muchas menos preguntas sobre características relevantes de los participantes y no participantes. Estos datos no proporcionaron estimaciones plausibles de diferencia simple mediante el método PSM (Correspondencia de puntaje de propensión) en comparación con las estimaciones de Jalan-Ravallion con respecto al mismo programa con mayor cantidad de datos. La explicación posible es que usar un instrumento de encuesta más ligero implicó que hubiera muchas diferencias no observables; en otras palabras, el supuesto de independencia condicional del PSM no fue válido. Dada la secuencia de las dos evaluaciones, se conocieron las variables clave omitidas en el último estudio, que principalmente se relacionaban con conexiones locales (como se evidencia en la pertenencia a asociaciones vecinales y en el tiempo vivido en el mismo barrio). Sin embargo, el instrumento de encuesta más ligero que usaron Ravallion et ál. (2005) tuvo la ventaja de que permitió realizar un seguimiento diacrónico de los mismos grupos familiares para formar un conjunto de datos de panel. Al parecer, Ravallion et ál. pudieron

⁴² De modo equivalente, miden el impacto por a través de la ganancia media entre los grupos familiares que no tienen tierra por vivir en una población que reúne los requisitos para el GB, menos la ganancia correspondiente a quienes tienen tierra.

resolver satisfactoriamente el problema del sesgo en el instrumento de encuesta más ligero realizando un seguimiento diacrónico de los grupos familiares, que les permitió distinguir los errores de correspondencia que surgían de los datos incompletos.

Esto ilustra un punto importante en el diseño de evaluaciones. Existe una tensión entre los recursos dedicados a recabar datos transversales para el pareo de diferencia simple y los dedicados a recabar datos longitudinales con instrumentos de encuesta más ligeros. Un factor importante a la hora de decidir qué método usar es cuánto sabemos *ex ante* acerca de los determinantes de la implementación del programa (tanto del lado de los administradores del programa como de los participantes). Si se puede implementar una sola encuesta que capte convincentemente estos determinantes, el método PSM funcionará bien. De lo contrario, es recomendable realizar al menos dos rondas de recopilación de datos y usar la DD, posiblemente en combinación con el PSM, como se expone a continuación.

Si bien los datos de panel no son esenciales cuando se calcula con el método DD, los datos de panel a nivel de los grupos familiares abren nuevas opciones para el análisis contrafactual de la distribución conjunta de resultados diacrónicos con el fin de dilucidar los impactos en la dinámica de la pobreza. Ravallion et ál. (1995) desarrollaron este enfoque con el fin de medir los impactos de los cambios en el gasto social en la distribución conjunta intertemporal del ingreso. En lugar de limitarse a medir el impacto sobre la pobreza (la distribución del ingreso marginal), los autores distinguen los impactos en el número de personas que escapa de la pobreza en el transcurso del tiempo (la función de “promoción” de una red de seguridad) de los impactos en el número que cae en la pobreza (la función de “protección”). Ravallion et ál. aplican este enfoque a una evaluación del impacto en las transiciones de la pobreza de la reforma en la red de seguridad social de Hungría. Se pueden encontrar otros ejemplos en Lokshin y Ravallion (2000) (sobre los impactos de los cambios en la red de

seguridad de Rusia durante una crisis financiera macroeconómica), Gaiha e Imai (2002) (sobre el Programa de Garantía de Empleo del estado hindú de Maharashtra) y van de Walle (2004) (sobre la evaluación de los resultados de la red de seguridad de Vietnam al abordar los impactos en los ingresos).

Los datos de panel también facilitan el uso de estimadores de regresión dinámica para la *DD*. Un ejemplo de este enfoque se puede encontrar en Jalan y Ravallion (2002), quienes identificaron los efectos de la dotación de infraestructura atrasada en un modelo dinámico de crecimiento del consumo utilizando un conjunto de datos de panel recabado en grupos familiares durante seis años. Su especificación econométrica es un ejemplo del modelo de efectos fijos no estacionario propuesto por Holtz-Eakin et ál. (1988), que tiene en cuenta los efectos geográficos e individuales latentes que pueden calcularse usando el Método Generalizado de Momentos (*Generalized Method of Moments*), que trata a los regresores que varían con el tiempo y al crecimiento retrasado del consumo como endógenos (usando retrasos suficientemente largos como variables instrumentales). Los autores encontraron aumentos significativos en el consumo a largo plazo a partir de las mejoras en las rutas de las zonas rurales.

Consideraciones acerca de los diseños de DD: Dos problemas clave han afectado a los estimadores de *DD* para evaluar los programas de lucha contra la pobreza en países en desarrollo. El primero de ellos es que, en la práctica, en el momento de realizar la encuesta inicial muchas veces no se sabe quién participará en el programa. Se debe realizar una estimación fundamentada para diseñar el muestreo para esta encuesta; puede resultar de ayuda conocer el diseño y el contexto del programa. En muchos casos, será necesario realizar un sobremuestreo de los tipos de unidades de observación que, por sus características, tengan más posibilidades de participar. Esto permite asegurar que se cubra de manera adecuada el grupo de tratamiento de la población y que se agrupe una cantidad suficiente de comparadores similares

sobre los cuales basarse. Con el tiempo, pueden surgir problemas si no se predice *ex ante* quién participará. Por ejemplo, Ravallion y Chen (2005) han diseñado su estudio de manera que el grupo de comparación se extraiga de poblaciones con muestreo aleatorio dentro de las mismas provincias pobres de la zona rural de China, donde se sabía que se encontrarían poblaciones de tratamiento (para un programa de desarrollo en zonas pobres). Sin embargo, los autores descubrieron más tarde que dentro de las provincias pobres había suficiente heterogeneidad y que muchas de las poblaciones seleccionadas para la comparación debían descartarse para asegurar el soporte común. La experiencia pasada indica que se necesita más esfuerzo para lograr un sobremuestreo de poblaciones relativamente pobres en países de escasos recursos.

El segundo problema es que el supuesto de *DD* del sesgo de selección invariable en el tiempo no se aplica a muchos programas de lucha contra la pobreza en países en desarrollo. Generalmente, en los programas de desarrollo para zonas pobres, se parte del supuesto de que éstas carecen de infraestructura y de otras dotaciones iniciales que, a su vez, generan un menor crecimiento y hacen que sigan siendo relativamente pobres. La *DD* por lo tanto será un estimador sesgado, debido a que los cambios posteriores en los resultados son una función de las condiciones iniciales que también influyeron en la asignación para el tratamiento. Por lo tanto, el sesgo de selección no es constante a través del tiempo. En la Figura 3, se explica este punto. Se grafican los resultados medios a través del tiempo, antes y después de la intervención. Los círculos levemente sombreados representan las medias observadas para las unidades de tratamiento, mientras que el círculo con trama es el contrafáctico en la fecha $t=1$. En el Panel (a), se muestra el sesgo de selección inicial, que surge del hecho de que el programa estaba destinado a zonas más pobres que las unidades de comparación (sombreadas de color oscuro). Esto no presenta un problema, siempre que el sesgo sea invariable en el tiempo, como en el panel (b). Sin embargo, cuando los atributos en los que se basa la asignación del programa también influyen en

las perspectivas de crecimiento posteriores, se obtiene un sesgo por defecto en el estimador de *DD*, como en el panel (c).

Dos ejemplos extraídos de evaluaciones reales ilustran el problema. Jalan y Ravallion (1998) muestran que los proyectos de desarrollo en zonas pobres de la zona rural de China estaban destinados a zonas con deficiencias en la infraestructura y que estas características fueron la causa de un menor crecimiento. Cabe suponer que las zonas con una infraestructura pobre tenían menos posibilidades de aprovechar las oportunidades generadas por el crecimiento económico de China. Jalan y Ravallion muestran que, en este caso, existe un importante sesgo en los estimadores de *DD*, debido a que los cambios a través del tiempo son una función de las condiciones iniciales (a través de un modelo de crecimiento endógeno) que también influyen en la implementación del programa. Al corregir este sesgo controlando las características de la zona que inicialmente se priorizaron al asignar los proyectos de desarrollo, los autores hallaron impactos significativos a largo plazo, a pesar de que ninguno de ellos había sido evidente en el estimador de *DD* estándar.

El segundo ejemplo está basado en el estudio de Pitt y Khandker (1998) del Grameen Bank. Según mi interpretación del método Pitt-Khandker para la evaluación de los impactos del crédito del GB, queda bastante claro que los autores presuponen que los retornos por poseer tierras son independientes de la elegibilidad del nivel de la población para acceder al GB. Surgirá un sesgo si el GB tiende a seleccionar poblaciones con retornos a la tierra inusualmente altos o bajos. Suena lógico que los retornos a la tierra sean menores en las poblaciones seleccionadas para el GB, lo que podría ser la principal causa de su pobreza. Además los bajos retornos a la tierra también podrían sugerir al GB que tales poblaciones presentan una ventaja comparativa en las actividades no agrícolas gracias al crédito del GB. Por consiguiente, el método Pitt-Khandker estaría sobreestimando el impacto del Grameen Bank.

De estas observaciones se puede concluir que es muy importante controlar la heterogeneidad inicial para que las estimaciones de DD sean creíbles. Una aparente medida correctiva es utilizar PSM para seleccionar el grupo de comparación inicial. Esto casi siempre reducirá el sesgo en las estimaciones de DD. En un ejemplo proveniente del contexto de programas de desarrollo en zonas pobres, Ravallion y Chen (2005) primero utilizaron PSM para eliminar la heterogeneidad inicial entre las poblaciones a quienes se destinaba el programa y las poblaciones del grupo de comparación, antes de aplicar la DD utilizando observaciones longitudinales para ambos grupos. Si los grupos de comparación inicial son relevantes, también pueden ayudar a reducir el sesgo en los estudios de DD (Galasso y Ravallion, 2004). El método de DD también se puede combinar con un diseño de discontinuidad (Jacob y Lefgren, 2004).

Estas observaciones indican importantes sinergias entre mejores datos y métodos para establecer comparaciones de diferencia simple (por un lado) y de doble diferencia (por el otro). Las observaciones longitudinales pueden ayudar a reducir el sesgo en las comparaciones de diferencia simple (eliminando la invariabilidad en el tiempo del sesgo de selección). Los intentos exitosos por eliminar la heterogeneidad en los datos iniciales, tal como puede realizarse por medio de PSM, también pueden reducir el sesgo en los estimadores DD.

¿Qué sucede si los datos iniciales no están disponibles? Generalmente, los programas de lucha contra la pobreza en los países en desarrollo se deben implementar rápidamente debido a una crisis macroeconómica o agroclimática, y no es posible retrasar la operación para realizar un estudio inicial. (No hace falta aclarar que tampoco está dentro de las opciones realizar una aleatorización). Aun así, en determinadas condiciones, es posible identificar los impactos en función de los resultados de los participantes, ante la ausencia del programa luego del mismo, y no antes. Para saber qué está en juego, se debe tener en cuenta que el supuesto de identificación clave en todos los estudios de doble diferencia es que el sesgo de selección del programa es

aditivamente separable de los resultados e invariable en el tiempo. En la implementación estándar descrita anteriormente en esta sección, la fecha 0 precede a la intervención y la *DD* proporciona la media de ganancias actuales para los participantes en la fecha 1. Supongamos, en cambio, que el programa está funcionando en la fecha 0. La posibilidad de identificación surge de que algunos participantes de la fecha 0 luego abandonan el programa. El estimador de triple diferencia (*DDD*) propuesto por Ravallion et ál. (2005) es la diferencia entre las dobles diferencias correspondientes a las personas que permanecen en el programa y las que abandonan. Ravallion et ál. muestran que su estimador *DDD* identifica en forma coherente la media de ganancias para los participantes en la fecha 1 (*TT*) si se cumplen dos condiciones: (i) no existe sesgo de selección en función de quién abandona el programa; (ii) no existen ganancias actuales para los que no participan. También muestran que una tercera encuesta permite evaluar en forma conjunta estas dos condiciones. Si se cumplen estas condiciones y no existe sesgo de selección en el período 2, no debería haber diferencias en la estimación de las ganancias para los participantes del período 1, ya sea que abandonen o no en el período 2.

Al aplicar el enfoque descrito anteriormente, Ravallion et ál. (2005) analizan qué sucede con los ingresos de los participantes cuando abandonan el programa *Trabajar* de Argentina en comparación con los ingresos de los participantes que siguen en el mismo, luego de registrar en cifras netas los cambios económicos, tal como lo revela un grupo de comparación pareado de no participantes. Los autores encuentran que se da una sustitución parcial de los ingresos; una cuarta parte del salario del programa *Trabajar* dentro de los seis meses de haberlo dejado, llegando a la mitad a los 12 meses. Por lo tanto, encuentran evidencias de una “caída de Ashenfelter” posterior

al programa; es decir, los ingresos disminuyen abruptamente por los recortes económicos, pero luego se recuperan.⁴³

Supongamos, en cambio, que no se cuenta con un grupo de comparación de no participantes. Se calcula la *DD* para las personas que permanecen versus las personas que abandonan (es decir, las ganancias a través del tiempo de las personas que permanecen, menos las ganancias de las que abandonan). Es indiscutible que esto sólo proporciona una estimación de las ganancias actuales de los participantes si los cambios contrafactuales a través del tiempo son los mismos para los que abandonan el programa y para los que permanecen. Uno podría esperar que, de no existir el programa, los que se quedan sean personas con menos posibilidades de obtener ganancias más a largo plazo que las personas que abandonan. Por lo tanto, la *DD* simple para las personas que permanecen versus las que abandonan subestimaré el impacto del programa. Ravallion et ál. hallaron que, en su contexto específico, la *DD* para los que permanecieron en el programa en comparación con los que lo dejaron (sin tener en cuenta los que nunca participaron) tuvo una aproximación bastante buena al estimador *DDD*. No obstante, es posible que esto no sea así en otras aplicaciones.

8. Flexibilización de la exogeneidad condicional

Aquí se abordarán los métodos que flexibilizan el supuesto de exogeneidad de OLS o PSM y que son resistentes al sesgo de selección variable en el tiempo, a diferencia de *DD*. Estos métodos tienen supuestos de identificación diferentes a los de los métodos anteriores.

Igualmente, se trata de supuestos que también pueden cuestionarse.

⁴³ La “caída de Ashenfelter” hace referencia al sesgo que se produce al utilizar la *DD* para inferir los impactos a largo plazo de los programas de capacitación, que pueden surgir cuando existe una caída en los ingresos antes del programa (según se describe en Ashenfelter, 1978).

Variables instrumentales: Volviendo al debate de la sección 2, supongamos que la implementación del programa depende de una variable instrumental (IV), Z , además de X :

$$T_i = \gamma Z_i + X_i \delta + v_i \quad (11)$$

(Luego se analizará de dónde puede provenir esta función). Para simplificar la explicación, nos centraremos en la especificación de impacto común (sección 2). El lector recordará que ésta es:

$$Y_i = ATE.T_i + X_i \beta^C + \mu_i^C \quad (5)$$

Si Z_i y X_i son exógenas, el sesgo de selección ($E(\mu^C | X, T) \neq 0$), entonces v_i y μ_i^C están potencialmente correlacionadas. La ecuación simplificada de los resultados es:

$$Y_i = \pi Z_i + X_i (\beta^C + ATE.\delta) + \mu_i \quad (12)$$

donde $\pi = ATE\gamma$ y $\mu_i = ATE v_i + \mu_i^C$. Si existe un Estimador de variables instrumentales (IVE, por su sigla en inglés) para el impacto medio, es $\hat{\pi}_{OLS} / \hat{\gamma}_{OLS}$ (en notación sencilla). Además de la exogeneidad de Z_i y de X_i , los supuestos clave para que $\hat{\pi}_{OLS} / \hat{\gamma}_{OLS}$ arroje una estimación coherente del impacto medio son que Z_i influye en la implementación ($\gamma \neq 0$, lo que asegura la existencia del IVE) y que Z_i no sea un elemento del vector de controles, X_i (lo que permite identificar π en (12) independientemente de β^C). La última condición se denomina “restricción de exclusión” (ya que Z_i se excluye de (5)). Si se cumplen estos supuestos, el IVE identifica el impacto medio del programa que es atribuible al instrumento y que acentúa el sesgo de selección. Una variante para este método es volver a escribir (11) como un modelo de respuesta binario no lineal (por ejemplo Probit o Logit) y utilizar el puntaje de propensión previsto como IV para la implementación del programa.⁴⁴

⁴⁴ Este estimador se analiza en Wooldridge (2002, Capítulo 18).

¿Cómo se compara el IVE con otros métodos? Al igual que los anteriores métodos NX, el IVE requiere un supuesto de independencia condicional no comprobable, a pesar de que se trata de un supuesto diferente al de PSM u OLS. En el caso del IVE, el supuesto no comprobable es la restricción de exclusión.⁴⁵ Sin embargo, se debe tener en cuenta que este supuesto no es estrictamente necesario cuando se utiliza una regresión de respuesta binaria no lineal para la primera etapa, a diferencia del modelo de probabilidad lineal de (11). Por lo tanto el modelo se identifica a partir de la no linealidad en la regresión de primera etapa. En la práctica, se prefiere ampliamente tener una estrategia de identificación que sea sólida, en lugar de utilizar una regresión lineal de primera etapa. Esto es un tema de debate, ya que la identificación a partir de la no linealidad sigue siendo identificación. Sin embargo, es preocupante cuando una identificación se basa en un supuesto un tanto *ad hoc* acerca de la distribución de un término de error. Para evitar esto se requiere una justificación para excluir Z_i de (5). Más adelante, se volverá sobre este tema.

También existen similitudes. Al igual que con OLS, la validez de las inferencias causales para el IVE (paramétrico) se basan principalmente en supuestos de forma funcional *ad hoc* para la regresión de resultados. También se debe tener en cuenta que la ecuación (11) de la primera etapa repite la primera etapa del método PSM. No obstante, posiblemente el IVE requiere menos de nuestra capacidad de modelar la asignación del programa que PSM; si bien la variable instrumental Z debe ser un predictor significativo para la participación, uno generalmente no se preocupa tanto por un R^2 bajo en la ecuación de la primera etapa para el IVE, como por el modelo utilizado para estimar los puntajes de propensión para parear o volver a realizar una ponderación.

⁴⁵ Si Z es un vector (con más de una variable) el modelo se sobreidentifica y se puede probar si todas menos una de las IV son significativas al sumarse a la ecuación principal de interés. Sin embargo, aún se debe dejar una IV y por lo tanto la restricción de exclusión no es comprobable.

Cabe destacar, además, que el IVE sólo identifica el efecto para un determinado subgrupo de población, concretamente los que fueron inducidos a participar en el programa por el instrumento. Naturalmente, la IV sólo puede revelar la variación exógena en la implementación del programa sólo para ese subgrupo. La ganancia resultante para el subgrupo inducido a cambiarse por la IV en algunos casos se denomina “efecto medio del tratamiento local” (LATE, por su sigla en inglés) (Imbens y Angrist, 1994). Este subgrupo generalmente no se identifica de manera explícita, por lo tanto en la práctica sigue quedando poco claro para quién exactamente se identificó el impacto medio.

El enfoque de función de control mencionado en la sección 5 proporciona además un método para abordar la endogeneidad. Al agregar una función de control adecuada (o “residual generalizada”) a la regresión de resultado se puede eliminar el problemático sesgo de selección de características no observables.⁴⁶ En general, el enfoque de CF debe proporcionar resultados similares al IVE. De hecho, las dos estimaciones son formalmente idénticas para una regresión de primera etapa lineal (como en la ecuación 11), debido a que luego el enfoque de control de función equivale a aplicar el OLS en (5) aumentados para incluir $\hat{v}_i = T_i - \hat{\gamma} Z_i$ como un regresor adicional (Hausman, 1978). Esta CF elimina la fuente del sesgo de selección, que se origina en el hecho de que $Cov(v_i, \mu_i^C) \neq 0$.

La restricción de exclusión: Se trata del talón de Aquiles del IVE en la práctica.

Hasta hace poco tiempo, el supuesto era apenas comentado en documentos que utilizaban el IVE (la opción de las IV en algunos casos incluso quedaba relegada a una nota al pie de una tabla de resultados del IVE, y prácticamente no volvía a mencionarse). No obstante, se pueden generar sesgos potencialmente importantes si la restricción no es válida. Se debe recordar que Glazeman

⁴⁶ Todd (2006) ofrece un interesante estudio global sobre estos métodos.

et ál. (2003) hallaron que este tipo de método para corregir el sesgo de selección en realidad tendía a aumentarlo, cuando se lo comparaba con resultados experimentales de los mismos programas. Los autores creen que esto se debe a restricciones de exclusión no válidas.

Pero en la actualidad han surgido estándares y se suele cuestionar la validez de la restricción de exclusión al analizar las evaluaciones de IVE en la práctica. Este cuestionamiento por lo general consiste en proponer algún modelo teórico alternativo para los resultados. Por ejemplo, consideremos el problema de identificar el impacto sobre los salarios de un programa de capacitación asignado en forma individual. Según la literatura anterior sobre economía laboral, se podrían utilizar las características del hogar al que pertenece cada persona como IV para la participación en el programa. Estas características influyen en la participación en el programa, pero los empleadores tienen pocas posibilidades de observarlas directamente. A partir de esto, se sostiene que las mismas no deberían afectar los salarios según se participe o no en el programa (otras variables de control observables, como por ejemplo, la edad y la educación del trabajador individual tampoco deberían influir). Sin embargo, al menos en algunas de estas IV potenciales, esta restricción de exclusión es cuestionable cuando hay efectos derivados, relevantes para la productividad, dentro de los hogares. Por ejemplo, se señaló que, en los países en desarrollo, la presencia de una persona alfabetizada en el hogar puede tener un efecto positivo en la productividad laboral de un analfabeto. Esto se encuentra fundamentado teóricamente y mediante evidencias (para la zona rural de Bangladesh) en Basu et ál. (2002).

¿Dónde se puede encontrar una IV? Existen fundamentalmente dos fuentes: las características del diseño experimental y los argumentos teóricos acerca de los determinantes de la implementación del programa y los resultados. A continuación se las considera por separado.

Diseños parcialmente aleatorios como fuente de variables instrumentales: Tal como se mencionó en la sección 4, en los experimentos sociales muchas veces sucede que las personas

seleccionadas al azar para el programa no desean participar. La asignación al azar es una elección natural para una IV en este caso. Aquí la restricción de exclusión es posible, es decir que el hecho de estar asignado al azar al programa sólo afecta los resultados a través de la participación real en el mismo.⁴⁷

En el experimento MTO, antes mencionado, puede encontrarse un ejemplo de este enfoque en la corrección del sesgo en diseños aleatorios. Familias estadounidenses seleccionadas al azar que vivían en zonas urbanas deprimidas, recibieron vales para comprar viviendas en zonas mejores. Naturalmente, no todas las personas a las que se les ofreció esta posibilidad la aceptaron. La diferencia en los resultados (como por ejemplo, las tasas de deserción escolar) sólo refleja el alcance del efecto externo (vecindad) si se corrige la participación endógena mediante una asignación al azar como IV (Katz et ál., 2001).

Se puede encontrar un ejemplo que corresponde a un país en desarrollo en el experimento *Proempleo*. Debe recordarse que éste incluyó un componente de capacitación que se asignó al azar. Bajo los supuestos de una participación del 100% o de una falta de cumplimiento aleatoria, ni el empleo ni los ingresos de los que recibieron la capacitación fueron significativamente diferentes a los del grupo de control 18 meses después del inicio del experimento.⁴⁸ Sin embargo, algunos a los que se les asignó el componente de capacitación no estaban interesados en él, y este proceso de selección estuvo correlacionado con los resultados de la capacitación. Se reflejó un impacto de la capacitación en las personas con educación secundaria, pero sólo cuando los autores corrigieron el sesgo de cumplimiento utilizando la asignación como IV para el tratamiento (Galasso et ál., 2004).

⁴⁷ Para obtener una descripción completa de las condiciones teóricas en las que un IVE proporciona el impacto medio de un programa, ver Angrist et ál. (1996). También se puede encontrar un análisis del tema en Dubin y Rivers (1997).

⁴⁸ El subsidio salarial incluido en el experimento *Proempleo* tuvo un impacto significativo en el empleo, pero no en los ingresos del momento, aunque es posible que los ingresos futuros hayan sido mayores. Ver Galasso et ál. (2004) para leer más sobre el tema.

El debate anterior se centró en el uso de la asignación al azar como IV para el tratamiento, dado el cumplimiento selectivo. Esta idea se puede extender al uso de la aleatorización en la identificación de los modelos económicos de resultados o al uso de comportamientos para determinar los resultados. Este tema volverá a tratarse en la sección 9.

Fuentes no experimentales de variables instrumentales: En la literatura de economía laboral que ha estimado regresiones de salarios con opción endógena de ocupación (o participación en la fuerza laboral), se encuentra una fuente común de IV al modelar el problema de opción ocupacional. Por lo tanto, se postula que existen variables que influyen en los costos de elección ocupacional, pero no en los ingresos de esa elección. Existe abundante literatura sobre tales aplicaciones del IVE y los estimadores relacionados.⁴⁹ En este caso, nos centraremos en las aplicaciones para evaluar los programas de lucha contra la pobreza. Las fuentes más comunes de variables instrumentales en este contexto fueron la ubicación geográfica de los programas, las variables políticas y las discontinuidades generadas por el diseño del programa.

El aspecto geográfico de la implementación del programa ha sido utilizado para la identificación en varios estudios. Se considerarán dos ejemplos. El primero es de Ravallion y Wodon (2000), quienes deseaban probar la tan difundida idea de que el trabajo de los menores interfiere con la escolaridad y, a largo plazo, perpetúa la pobreza. Utilizaron un subsidio que se estaba otorgando para la matrícula de una escuela en la zona rural de Bangladesh (el Programa *Alimentos por educación [Food-for-Education Program]*) como la fuente de cambio en el costo de la escolaridad en su modelo de escolaridad y empleo de menores. Para abordar la endogeneidad de la implementación del programa a nivel personal, utilizaron como IV la implementación anterior a nivel de la población. La preocupación aquí es la posibilidad de que la implementación en la población esté correlacionada con factores geográficos relevantes para los

⁴⁹ Se puede consultar un excelente estudio global sobre el tema en Heckman et ál. (1999).

resultados. En base a información externa sobre las normas administrativas de asignación, Ravallion y Wodon proporcionan pruebas de exogeneidad que apoyan su estrategia de identificación, a pesar de que, en última instancia, ésta depende de una restricción de exclusión no comprobable y/o no linealidad para la identificación. Los resultados indican que el subsidio aumentó la escolaridad en una proporción mucho mayor que la proporción en que redujo el trabajo de menores. Según parece, los efectos de sustitución ayudaron a proteger los ingresos actuales provenientes de la mayor asistencia escolar inducida por el subsidio.

Se puede encontrar un segundo ejemplo de este enfoque en Attanasio y Vera-Hernandez (2004), quienes estudiaron los impactos de un importante programa de nutrición en la zona rural de Colombia, que proporcionó alimentos y atención infantil a través de centros comunitarios locales. Algunas personas utilizaron estos recursos y otras no, y es bastante justificado suponer que, en este contexto, el uso es endógeno a los resultados. Para solucionar este problema, Attanasio y Vera-Hernandez utilizaron la distancia desde un hogar hasta el centro comunitario como IV para la asistencia a dicho centro. Los autores también se ocupan de las objeciones que pueden hacerse en contra de la restricción de exclusión.⁵⁰ La distancia en sí podría ser endógena a todas las elecciones de ubicación de los hogares y los centros comunitarios. Entre las justificaciones que mencionan los autores para su elección de IV, destacan que los encuestados que se mudaron hace poco tiempo no mencionan la necesidad de mudarse más cerca de centros comunitarios como una razón para elegir la ubicación (aun cuando era una de las opciones). Destacan también que si sus resultados hubiesen estado realmente determinados por la endogeneidad de su IV, encontrarían efectos (artificiales) en variables que no deberían afectarse,

⁵⁰ Al igual que en el ejemplo de Ravallion-Wodon, en este caso se cumple con mayor facilidad el otro requisito para que una IV sea válida, es decir que esté correlacionada con el tratamiento.

como el peso al nacer. Sin embargo, no hallaron dichos efectos, lo que apoya la elección de su IV.

Las características políticas de las zonas geográficas fueron otra fuente de instrumentos. Conocer la economía política del lugar donde se implementa el programa puede ayudar a identificar los impactos. Por ejemplo, Besley y Case (2000) utilizan la presencia de mujeres en los parlamentos de los estados (dentro de los EE.UU.) como IV para el seguro de accidentes laborales al estimar los impactos de la indemnización en los salarios y el empleo. Los autores suponen que las legisladoras están a favor de la indemnización de los empleados pero que esto no tiene un efecto independiente en el mercado laboral. La última condición no se cumpliría si una mayor incidencia de mujeres en el parlamento de determinado estado reflejara factores sociales latentes que generaran una mayor participación femenina en la fuerza laboral en general, con repercusiones en los resultados totales del mercado laboral, tanto de mujeres como de hombres.

Para ofrecer otro ejemplo, en la evaluación de fondos sociales financiados por el Banco en Perú, Paxson y Schady (2002) utilizaron el grado en que las elecciones recientes dejaron de apoyar al gobierno como IV para la asignación geográfica de los gastos del programa en la explicación de los resultados escolares. Su idea era que la asignación geográfica de los gastos del fondo social se podría haber utilizado en parte para “volver a comprar” a los votantes que dejaron de apoyar al gobierno en las últimas elecciones. (La regresión de primera etapa concuerda con la hipótesis que presentaron). También se debe suponer que el hecho de que un sector haya dejado de apoyar al gobierno en las últimas elecciones no está correlacionado con factores latentes que influyen en la escolaridad. Se halló que la variación en los gastos atribuida a esta IV aumentó significativamente la tasa de escolaridad.

El tercer grupo de ejemplos pone de manifiesto las discontinuidades en el diseño del programa, de acuerdo con lo analizado en la sección 6. En este caso, el LATE está próximo a un corte en la elegibilidad del programa. Se puede encontrar un ejemplo de este enfoque en Angrist y Lavy (1999), quienes evaluaron el impacto del tamaño de las clases en los niveles académicos obtenidos en Israel. A los fines de la identificación, utilizaron el hecho de que (en Israel) se asignó un maestro adicional cuando el tamaño de la clase superaba los 40 alumnos. Sin embargo, no existen razones posibles para determinar por qué este punto de corte en el tamaño de la clase podría tener un efecto independiente en los niveles obtenidos, y por consiguiente justificar la restricción de exclusión. Los autores encuentran considerables beneficios en los tamaños de clase más reducidos, que no resultaban evidentes al utilizar OLS.

Otro ejemplo se encuentra en el estudio de Duflo (2003) sobre los impactos de las pensiones por ancianidad de Sudáfrica en los indicadores antropométricos infantiles. Las mujeres sólo son elegibles para estas pensiones a los 60 años de edad, mientras que los hombres son elegibles a los 65. Es muy poco probable que haya una discontinuidad en los resultados (condicionales del tratamiento) a estas edades críticas. Siguiendo a Case y Deaton (1998), Duflo utilizó la elegibilidad como IV para la recepción de pensiones por ancianidad en sus regresiones para las variables de resultado antropométricas. Halló que las pensiones que reciben las mujeres mejoran el estado nutricional de las niñas pero no de los niños, mientras que las pensiones que reciben los hombres no tienen efectos sobre los resultados, ni para los niños ni para las niñas.

Una vez más, esto supone que se conoce la elegibilidad, pero no siempre es así. Además, la elegibilidad para los programas de lucha contra la pobreza por lo general se basa en criterios de pobreza, que también son las variables de resultado relevantes. Por lo tanto, al estimar quién es elegible (para construir la IV) se debe cuidar no partir de supuestos que prejuzguen los impactos del programa.

Se pueden realizar dos observaciones acerca de la relación existente entre estos métodos y los diseños de discontinuidad analizados en la sección 6, a través de lo cual se realiza una comparación de diferencia simple de las medias a cada lado del punto de corte. En primer lugar, y de modo similar al problema mencionado anteriormente sobre el cumplimiento selectivo en un diseño aleatorio, el uso de la discontinuidad en la regla de elegibilidad como una IV para la implementación real del programa, puede abordar las preocupaciones acerca del cumplimiento selectivo de esas reglas. Este tema se analiza con más detalle en Battistin y Rettore (2002). En segundo lugar, estos métodos de IV por lo general no arrojan los mismos resultados que los diseños de discontinuidad abordados en la sección 6. En Hahn et ál., se establecen condiciones específicas para la equivalencia entre estos dos métodos. Las principales condiciones son que las medias utilizadas en la comparación de diferencia simple se calculen a través de las correspondientes ponderaciones *kernel* y que el estimador IVE se aplique a una submuestra específica, en las proximidades del punto de corte de elegibilidad.

Tal como lo indican estos ejemplos, la justificación de un IVE en última instancia debe basarse en las fuentes de información que se encuentran fuera de los límites del análisis cuantitativo. Estas fuentes pueden incluir argumentos teóricos, de sentido común o argumentos empíricos basados en diferentes tipos de datos, incluidos datos cualitativos, como por ejemplo los que se basan en el conocimiento de cómo funciona el programa en la práctica.

Límites sobre el impacto: En la práctica, el IVE muchas veces arroja estimaciones de impacto poco posibles (demasiado pequeñas o demasiado grandes). Cabría sospechar que el motivo es una violación de la restricción de exclusión. ¿Pero cómo se podrían emitir juicios acerca de este tema de manera más científica? Si es posible descartar *a priori* determinados valores de *Y*, se podrían establecer límites factibles a las estimaciones de impacto (siguiendo un

enfoque presentado por Manski, 1990). Esto resulta fácil si la variable de resultado es “pobre” versus “no pobre” (o algún otro resultado binario). Entonces $0 \leq TT \leq E(Y^T|T=1)(\leq 1)$ y:⁵¹

$$\begin{aligned} & (E[Y^T|T=1]-1)\Pr(T=1) - E[Y^C|T=0]\Pr(T=0) \\ & \leq ATE \leq \\ & (1-E[Y^C|T=0])\Pr(T=0) + E[Y^T|T=1]\Pr(T=1) \end{aligned}$$

La separación entre estos límites dependerá de los datos concretos del contexto. Los límites no serán de mucha utilidad en el caso (común) de las variables de resultado continuas.

Se sugiere otro enfoque para calcular los límites de las estimaciones de impacto en Altonji, Elder y Taber (AET) (2005a,b) en su estudio del efecto de asistir a un colegio católico sobre la escolaridad en los EE.UU. Los autores reconocen el posible sesgo en las estimaciones de OLS de esta relación (posiblemente sobrestimar el impacto real), pero cuestionan además las restricciones de exclusión utilizadas en las estimaciones de IV anteriores. Es necesario recordar que OLS toma el supuesto de que las características no observables que afectan los resultados no están correlacionadas con la implementación del programa. AET estudian las repercusiones del supuesto de alternativa extrema: que las características no observables de los resultados tienen el mismo efecto en la implementación que en el índice de las observables (el término $X_i\beta^C$ en (5)). Dicho de otro modo, se supone que la selección de las características no observables es tan importante como la de las observables.⁵² Para implementar este supuesto se debe restringir el coeficiente de correlación entre los términos de error de las ecuaciones correspondientes a los resultados y a la participación (μ^C en (5) y ν en (11)) a un valor dado por el coeficiente de

⁵¹ El límite menor para *ATE* se calcula estableciendo $E[Y^T|T=0]=0$ y $E[Y^C|T=1]=1$; el límite mayor: $E[Y^T|T=0]=1$, $E[Y^C|T=1]=0$.

⁵² Altonji et ál. (2005a) hablan de las condiciones para que esto se cumpla. Sin embargo, tal como señalan, no se espera que estas condiciones se cumplan en la práctica. Su estimador proporciona un límite para la estimación real, en lugar de una estimación de punto alternativa.

regresión de la función de puntaje correspondiente a las características observables en la ecuación de participación ($X_i\delta$ en la ecuación (11) con $\gamma = 0$) en la correspondiente función de puntaje para los resultados ($X_i\beta^C$).

AET sostienen que su estimador proporciona un límite menor para el impacto real cuando este último es positivo. Esto se basa en el supuesto (razonable si se considera *a priori*) de que el término de error en la ecuación de resultados incluye al menos algunos factores que verdaderamente no están relacionados con la participación. La estimación de OLS proporciona un límite mayor. Por lo tanto, el estimador de AET indica la sensibilidad de OLS a cualquier sesgo de selección basado en las características no observables. Por ejemplo, Altonji et ál. (2005a) hallaron que asistir a un colegio católico tiene un impacto de ocho puntos porcentuales en la tasa de egreso del colegio secundario cuando se supone la exogeneidad, pero que este impacto disminuye a cinco puntos al utilizar su estimador. Esto sugiere también una prueba de especificaciones del IVE. Se podría cuestionar un IVE que se encuentre fuera del intervalo entre los estimadores de AET y OLS.⁵³

9. Conocimientos que se obtienen de las evaluaciones

Hasta el momento, el centro de atención estuvo en la pregunta sobre la “validez interna”: ¿el diseño de la evaluación nos permite obtener una estimación confiable de los resultados contrafactuales en el contexto específico? Éste ha sido el objetivo principal de la literatura hasta la fecha. No obstante, existen intereses de igual importancia relacionados con los conocimientos que se pueden obtener de una evaluación de impacto más allá de su contexto específico. Esta sección está dedicada a la pregunta sobre la “validez externa”; es decir, si los resultados de

⁵³ Altonji et ál. (2005b) muestran la manera en que su método se puede utilizar también para evaluar el sesgo potencial en IVE debido a una restricción de exclusión no válida.

determinadas evaluaciones se pueden aplicar en otros contextos (lugares y/o fechas) y qué se puede aprender de la investigación evaluativa para aplicar en el conocimiento del desarrollo y las políticas futuras.

¿Los sesgos de publicación impiden que se puedan obtener conocimientos de las evaluaciones? La formulación de las políticas de desarrollo se basa en el conocimiento acumulado proveniente de evaluaciones publicadas. Por lo tanto, los procesos de publicación y los incentivos para los investigadores son relevantes en el éxito de la lucha contra la pobreza y en el logro de otros objetivos de desarrollo.

No resultaría sorprendente encontrar que es más difícil publicar un trabajo que presenta impactos no esperados o ambiguos, cuando se los compara con teorías aceptadas o con evidencias anteriores. Los revisores y editores pueden aplicar estándares diferentes al juzgar los datos y los métodos en función de si creen en los resultados de manera *a priori*. En la medida en que generalmente se esperan impactos de los programas de lucha contra la pobreza (que es probablemente el principal motivo por el que existen estos programas), esto significará que nuestro conocimiento está sesgado a favor de los impactos positivos. Al analizar un nuevo tipo de programa, los resultados de los primeros estudios sentarán precedentes para la evaluación de trabajos posteriores. El hecho de que se evalúe incorrectamente la distribución real inicial de los impactos puede ocasionar una posterior distorsión temporal en la distribución conocida. Sin duda, estos sesgos pueden afectar la producción de la investigación evaluativa, así como también las publicaciones. Los investigadores pueden esforzarse por obtener hallazgos positivos para aumentar las posibilidades de publicar su trabajo. No existen dudas de que los sesgos importantes (en cualquier dirección) saldrán a la luz en algún momento, pero esto tardará un tiempo.

Se trata, en buena parte, de conjeturas propias. Una evaluación precisa requiere alguna forma de inferir la distribución contrafactual de los impactos, ante la ausencia de sesgos de publicación. Evidentemente esto es muy difícil en la mayoría de los casos. Sin embargo, queda por lo menos un sector de la investigación evaluativa donde el sesgo de publicación es poco probable. Se trata de los estudios de replicación que han comparado los resultados NX con hallazgos experimentales para los mismos programas (como en el metaestudio de Glazerman et ál. 2003, sobre programas laborales en países desarrollados). Comparar la distribución de las estimaciones de impacto publicadas provenientes de estudios NX (sin replicación) con un contrafáctico extraído de los estudios de replicación del mismo tipo de programa, podría aportar datos relevantes en cuanto al grado de sesgo de publicación.

¿Es posible “escalar” los conocimientos obtenidos de una evaluación? El contexto de una intervención generalmente tiene gran peso sobre los resultados, por lo tanto pueden confundirse las inferencias al “escalar” una evaluación de impacto. (Estas cuestiones de “validez externa” se relacionan con evaluaciones experimentales y NX). Si se incluyen factores contextuales, puede resultar difícil realizar generalizaciones significativas para escalar y replicar pruebas. El mismo programa funciona a la perfección en una población pero falla irremediamente en otra. Esto se ve claramente en Galasso y Ravallion (2005), quienes estudian el Programa *Alimentos por educación* de Bangladesh. El programa llegó bien a los pobres en algunas poblaciones pero no en otras, a pesar de que estaban bastante cerca.

El punto clave aquí es que el contexto institucional de una intervención puede determinar en gran medida sus impactos. Pueden aparecer cuestiones de validez externa sobre las evaluaciones de impacto cuando es necesario que estén presentes algunas instituciones, incluso para facilitar los experimentos. Por ejemplo, cuando las pruebas aleatorias están ligadas a actividades de determinadas Organizaciones no gubernamentales (ONG, o NGO por sus siglas

en inglés) que son facilitadoras (como los casos citados por Duflo y Kremer, 2005), se teme que la misma intervención a escala nacional tenga un impacto muy diferente en lugares donde no estén estas organizaciones. Puede resultar de ayuda asegurar que las zonas del grupo de control cuenten con una ONG, pero aun así no se pueden descartar los efectos de interacción entre las actividades de la ONG y la intervención. En otras palabras, es posible que el efecto de la ONG no sea “aditivo” sino “multiplicativo”, de manera tal que la diferencia entre los resultados medidos para el tratamiento y para los grupos de control no revelen el impacto de la ausencia de la ONG.

Otro problema de validez externa es que, si bien los supuestos de equilibrio parcial pueden ser acertados para un proyecto piloto, los efectos de equilibrio general (a veces denominados efectos de “feedback” o “macro” efectos en la literatura de evaluación) pueden ser importantes al realizar el escalamiento a nivel nacional. Por ejemplo, una estimación del impacto de un subsidio educativo en la escolaridad, basado en una prueba aleatoria puede resultar engañosa al realizar un escalamiento, debido a que se alterará la estructura de los retornos de la escolaridad.⁵⁴ Otro ejemplo: un pequeño programa piloto de subsidio salarial como el implementado en el experimento *Proempleo* probablemente no tenga demasiado impacto en la tasa salarial del mercado, pero esto cambiará al escalar el programa. Una vez más, surgen problemas de validez externa como consecuencia de la especificidad contextual de las pruebas. Los resultados en el contexto de la prueba pueden variar en gran medida (en cualquier dirección) una vez que se escala la intervención, y los precios y salarios responden.

⁵⁴ Heckman et ál. (1998) demuestran que el análisis de equilibrio parcial puede sobreestimar en gran medida el impacto de un subsidio educativo una vez que se ajustan los salarios relativos. Por su parte, Lee (2005) encuentra una diferencia mucho menor entre los efectos de equilibrio general y parcial de un subsidio educativo en un modelo un tanto diferente.

Es evidente que los factores contextuales son cruciales para las políticas y el desarrollo de los programas. A riesgo de exagerar, se puede decir que en algunos contextos funcionará todo y en otros fallará todo. Generalmente, un factor clave para el éxito de un programa es la adaptación correcta al contexto institucional y socioeconómico donde se debe trabajar. Eso es lo que hace, en todo momento, el personal de un buen proyecto. Podrían basarse en los conocimientos provenientes de evaluaciones anteriores, pero éstos casi nunca son decisivos y pueden resultar realmente engañosos si se utilizan de manera mecánica.

Los impactos reales en el escalamiento también pueden diferir de los resultados de la prueba (sea aleatoria o no) debido a que la composición socioeconómica de la participación en el programa varía con el escalamiento. Ravallion (2004a) analiza cómo puede ocurrir esto y presenta resultados provenientes de varios estudios de caso, que sugieren que, luego del escalamiento, la incidencia de los beneficios del programa se torna más pro pobre. Los resultados de las pruebas pueden subestimar cuán pro pobre será un programa luego del escalamiento, ya que la economía política sugiere que los beneficios iniciales serán aprovechados primero por los no pobres (Lanjouw y Ravallion, 1999).

¿Qué determina el impacto? Estas cuestiones de validez externa indican la necesidad de complementar las herramientas de evaluación antes descritas con otros recursos de información que puedan ayudar a comprender los procesos que influyen en los resultados medidos.

Un enfoque es repetir la evaluación en diferentes contextos, según lo propuesto por Duflo y Kremer (2005). Se puede encontrar un ejemplo en el ya mencionado estudio de Galasso y Ravallion, en el que se evaluó el impacto del programa *Alimentos por educación* de Bangladesh en cada una de las 100 poblaciones de Bangladesh, y los resultados se correlacionaron con las características de esas poblaciones. Los autores hallaron que las diferencias en los resultados del programa podían explicarse, en parte, en términos de las características observables de las poblaciones, como por ejemplo el grado de desigualdad dentro de las mismas (en las que había más desigualdad fue más difícil llegar a los pobres a través del programa). Una forma de solucionar estas cuestiones es repetir las evaluaciones en diferentes contextos y a diferentes escalas. Sigue siendo tema de debate la posibilidad realizar una cantidad suficiente de pruebas (para cubrir la gran variación que se encuentra en la realidad). La escala que debe tener una

prueba aleatoria para probar un programa nacional de gran magnitud podría llegar a ser prohibitiva. No obstante, es una buena idea variar los contextos de las pruebas, siempre que sea viable.

Un enfoque alternativo es analizar en profundidad por qué un programa tiene (o no) impacto en un determinado contexto, como punto de partida para inferir si funcionaría en otro contexto. El diseño de evaluación más común identifica una cantidad bastante reducida de indicadores de “resultado final” e intenta evaluar el impacto del programa en función de estos indicadores. Sin embargo, en lugar de utilizar únicamente indicadores de resultado final, es posible estudiar también los impactos sobre determinados indicadores intermedios de comportamiento. Por ejemplo, las respuestas conductuales intertemporales de los participantes en los programas de lucha contra la pobreza son bastante relevantes para comprender sus impactos. Una evaluación de impacto de un programa de transferencias compensatorias de dinero a los agricultores mexicanos mostró que el dinero de las transferencias era invertido en parte, con efectos de segundo orden en los ingresos futuros (Sadoulet et ál., 2001). Del mismo modo, Ravallion y Chen (2005) hallaron que los participantes de un programa de desarrollo de zonas pobres de China ahorraron gran parte de los ingresos obtenidos del programa (según se estimó mediante el método de doble diferencia pareada descrito en la sección 7). Identificar las respuestas a través de los ahorros y la inversión ayuda a comprender los impactos actuales sobre el nivel de vida y las posibles ganancias futuras de asistencia social, más allá de la duración del proyecto. En lugar de analizar únicamente el indicador de asistencia social acordado, se recogen y analizan datos sobre una gran cantidad de indicadores intermedios relevantes para comprender los procesos que determinan los impactos.

Esto describe además un aspecto común en los estudios de evaluación; a saber, que el período del estudio casi nunca es mucho más largo que el período de desembolsos del programa. Sin embargo, parte del impacto sobre el nivel de vida de las personas puede producirse una vez

que haya finalizado el proyecto. Esto no necesariamente implica que las evaluaciones creíbles deban realizar un seguimiento de los impactos de la asistencia social durante períodos mucho más prolongados que los habituales, lo que pondría en juego cuestiones de credibilidad. Pero sí sugiere que las evaluaciones deben analizar detenidamente los impactos sobre los indicadores intermedios parciales de impactos más a largo plazo, aun cuando se dispone de buenas medidas del objetivo de asistencia social dentro del ciclo del proyecto. La elección de tales indicadores deberá estar fundamentada por las respuestas conductuales de los participantes al programa.

Para obtener conocimientos de una evaluación, por lo general es necesario basarse en gran medida en información externa a la evaluación. La investigación cualitativa (entrevistas intensivas con los participantes y los administradores) puede ser una fuente de información útil.⁵⁵ Un enfoque es utilizar métodos cualitativos para probar los supuestos de una intervención. Esto a veces se denomina “evaluación basada en la teoría” a pesar de que no sea un término ideal, debido a que las estrategias de identificación NX para los impactos medios generalmente están basadas en la teoría (tal como se explicó en la sección anterior). Weiss (2001) describe teóricamente este enfoque en el contexto de la evaluación de impactos de los programas comunitarios de lucha contra la pobreza. Se puede encontrar un ejemplo en una evaluación de fondos sociales (SF, por sus siglas en inglés) del Departamento de Evaluación de Operaciones del Banco Mundial, de acuerdo con lo resumido en Carvalho y White (2004). Si bien el objetivo general de un SF es reducir la pobreza, el estudio del OED intentó determinar si los SF funcionaban tal como lo pretendían las personas que los diseñaron. Por ejemplo, ¿participaron las comunidades locales? ¿Quién participó? ¿La clase alta local se “quedó” con el SF (como sostenían algunos críticos)? De acuerdo con Weiss (2001), la evaluación del OED identificó varios vínculos clave hipotetizados entre la intervención y los resultados, y además permitió

⁵⁵ Ver el análisis sobre “métodos combinados” en Rao y Woolcock (2003).

determinar si todos funcionaban. Por ejemplo, en uno de los estudios correspondiente a la evaluación que hizo el OED de los SF, Rao e Ibanez (2005) probaron el supuesto de que un SF funciona en respuesta a los subproyectos que proponen en forma conjunta las comunidades locales. En un SF de Jamaica, los autores hallaron que el proceso generalmente estaba dominado por la clase alta local.

En la práctica, es muy poco probable que todos los supuestos relevantes sean comprobables (incluidos los supuestos alternativos provenientes de teorías diferentes que podrían generar impactos similares). Tampoco es verdad que el proceso que determina el impacto de un programa se pueda dividir siempre en una serie de vínculos comprobables dentro de una cadena causal única. Puede haber formas de interacción y simultaneidad más complejas que no se prestan a este tipo de análisis. Por estos motivos, el enfoque denominado “evaluación basada en la teoría” no puede considerarse un sustituto serio para evaluar impactos en los resultados finales, con métodos (experimentales o NX) creíbles. Sí podría llegar a ser un complemento útil de dichas evaluaciones, para comprender mejor los impactos medidos.

Las bases de datos para el monitoreo del proyecto son una fuente importante de información, que no se aprovecha completamente. Con frecuencia los datos de monitoreo del proyecto y el sistema de información tienen contenido evaluativo sin demasiada importancia. Pero no siempre es así. Por ejemplo, la idea de combinar mapas de gastos con mapas de pobreza para evaluar rápidamente los resultados de un programa descentralizado de lucha contra la pobreza es un prometedor ejemplo de cómo, con costos módicos, se pueden aprovechar los datos de monitoreo estándar para ofrecer información sobre el funcionamiento del programa, y con una retroalimentación lo suficientemente rápida como para que el proyecto se pueda corregir sobre la marcha (Ravallion, 2001).

El experimento *Proempleo* es un ejemplo de que la información externa a la evaluación puede aportar importantes conocimientos para realizar un escalamiento. Recordemos que *Proempleo* asignó al azar vales para un subsidio salarial a personas (generalmente pobres) que se encontraban en un programa para desempleados y realizó un seguimiento de su posterior éxito para obtener un trabajo normal. Un grupo de control asignado al azar localizó el contrafáctico. Los resultados indican que hubo un impacto significativo del vale de subsidio salarial sobre el empleo. Pero cuando éstos se compararon con los datos administrativos centrales, complementados con entrevistas informales con las empresas de contratación, se halló que hubo muy poca aceptación del subsidio salarial por parte de las empresas (Galasso et ál., 2004). El esquema fue altamente rentable: el gobierno ahorró el 5% de su gasto en subsidios de desempleo para un desembolso por subsidios que representó sólo el 10% de ese ahorro.

Sin embargo, las comparaciones complementarias con otros datos revelaron que *Proempleo* no funcionó tal como se pretendió al ser diseñado. La mayor ganancia en cuanto a empleo para los participantes no surgió de la mayor demanda de empleo inducida por el subsidio salarial. El impacto surgió por los efectos laterales de la oferta: el vale tenía cierto privilegio entre los empleados; era como una “carta de presentación” que sólo unos pocos tenían (y en la zona nadie sabía cómo se realizaba la asignación).

Esto no podía ser revelado por la evaluación (aleatoria), sino que requería información complementaria. La otra perspectiva que se obtuvo acerca de cómo funcionó realmente *Proempleo* en el contexto de su prueba también tuvo repercusiones para el escalamiento, que puso más énfasis en informar mejor a los trabajadores pobres sobre cómo obtener un trabajo en lugar de otorgarles subsidios salariales.

Los efectos derivados también indican la importancia de comprender en profundidad cómo funciona un programa. Es común que haya impactos indirectos (o de “segunda vuelta”) en

los no participantes. Un programa para desempleados puede resultar en mayores ingresos para los no participantes. O bien un proyecto de mejora de ruta en determinada zona puede mejorar el acceso a otros lugares. En función de la importancia que se otorgue a estos efectos indirectos en la aplicación específica, es posible que haya que redefinir el “programa” para incluir los efectos derivados. O bien se podría combinar el tipo de evaluación que se analiza aquí con otras herramientas, como por ejemplo un modelo del mercado laboral para obtener otros beneficios.

La forma extrema de un efecto derivado es un programa que afecta todos los sectores de la economía. Las herramientas de evaluación analizadas en este capítulo son para programas asignados, pero tienen un rol poco evidente en los programas que abarcan todos los sectores de la economía, donde ningún proceso de asignación explícito es evidente, o si lo es, los efectos derivados probablemente sean dominantes. Si el programa que abarca todos los sectores de la economía se implementa en algunos países pero no en otros, el trabajo comparativo entre los países (tales como las regresiones de crecimiento) puede revelar impactos. Esa tarea de identificación por lo general resulta difícil, especialmente porque suelen existir factores latentes a nivel país que influyen en forma simultánea en los resultados y en si un país adopta la política en cuestión. Incluso cuando se acepta la estrategia de identificación, puede resultar realmente complejo aplicar los resultados generalizados de las regresiones entre países para fundamentar la formulación de políticas en cualquier otro país. También existen varios ejemplos prometedores de cómo se pueden combinar las herramientas de simulación para las políticas que se aplican a todos los sectores de la economía, como por ejemplo los modelos de Equilibrio General Computable, que se pueden combinar con los datos de encuestas al grupo familiar para evaluar los impactos sobre la pobreza y la desigualdad.⁵⁶ Estos métodos de simulación hacen que sea

⁵⁶ Ver, por ejemplo, Bourguignon et ál. (2003), y Chen y Ravallion (2004).

mucho más fácil atribuir impactos al cambio de política, pero esta ventaja tiene el costo de que es necesario trabajar sobre muchas más hipótesis acerca de cómo funciona la economía.

¿La evaluación responde a preguntas relevantes para la formulación de políticas? Es indiscutible que lo que se pretende lograr en última instancia con cualquier evaluación es obtener conocimientos para futuras políticas. Las prácticas de evaluación estándar pueden resultar poco informativas, cuando se las examina más detenidamente.

Un problema es la elección del contrafáctico. La formulación clásica del problema de evaluación analiza los impactos medios en las personas que reciben el programa, en relación con los resultados contrafactuales en ausencia del programa. No obstante, esto puede estar bastante alejado de las inquietudes de las autoridades responsables de la formulación de las políticas. Si bien la práctica común es utilizar los resultados en ausencia del programa como contrafáctico, generalmente la alternativa que prefieren los responsables de la formulación de políticas es destinar los mismos recursos a otro programa (posiblemente una versión diferente del mismo programa), en lugar de no hacer nada. El problema de la evaluación no cambia formalmente si pensamos en algún programa alternativo como contrafáctico. En principio, también se podría repetir el análisis del “contrafáctico de no hacer nada” para cada alternativa posible y comparar los resultados, aunque esto es difícil en la práctica. Es posible que un programa parezca funcionar bien al compararlo con la opción de no hacer nada, pero éste puede resultar deficiente si se lo compara con alguna posible alternativa.

Por ejemplo, mediante su evaluación del impacto de un programa para desempleados en India, Ravallion y Datt (1995) muestran que el programa redujo considerablemente la pobreza frente al contrafáctico de no aplicar ningún programa. No obstante, una vez que se tuvieron en cuenta los costos del programa (incluidos los ingresos previstos de los participantes en el programa de asistencia para desempleado), los autores concluyeron que el contrafáctico

alternativo de una asignación uniforme (sin destinatarios específicos) del mismo desembolso del presupuesto hubiera tenido mayor impacto sobre la pobreza.⁵⁷

Otro problema, que está más relacionado con los métodos utilizados para la evaluación, es si se han identificado los parámetros de impacto más relevantes desde el punto de vista de la cuestión de formulación de políticas. La formulación clásica del problema de evaluación se centra en los resultados medios, como por ejemplo la media de ingresos o consumo. Esto no suele ser lo adecuado para programas cuyo objetivo (más o menos) explícito es reducir la pobreza, más que promover el crecimiento económico *per se*. Sin embargo, tal como se aclaró en la sección 3, no existe nada que impida volver a interpretar la medida de resultado de manera que (2) proporcione el impacto del programa en un índice de recuento de la pobreza (% por debajo de la línea de pobreza). Al repetir el cálculo de impacto para varias “líneas de pobreza”, se puede trazar el impacto sobre la distribución acumulativa de los ingresos. Es posible hacer esto con las mismas herramientas, a pesar de que la práctica de evaluación haya estado centrada dentro de límites un tanto más estrechos.

Generalmente existe un interés por comprender mejor los impactos horizontales del programa; es decir, las diferencias de impactos en un determinado nivel de resultados contrafactuales, tal como lo se ve en la distribución conjunta de Y^T e Y^C . No es posible determinar esto a través de un experimento social, que sólo revela los resultados netos contrafactuales medios correspondientes a las personas tratadas. TT proporciona la media de ganancias netas de las pérdidas entre los participantes. En lugar de centrarse únicamente en las ganancias netas de los pobres (por ejemplo), uno se podría preguntar cuántos perdedores y ganadores hubo entre los pobres. En la sección 7, se presentó un ejemplo; a saber, el uso de datos de panel para estudiar los impactos de un programa de lucha contra la pobreza en la dinámica de

⁵⁷ Para obtener otro ejemplo del mismo resultado, ver Murgai y Ravallion (2005).

la pobreza. Algunas intervenciones pueden arrojar perdedores a pesar de que el impacto medio sea positivo, y naturalmente los responsables de la formulación de las políticas desearán obtener información acerca de esos perdedores, al igual que de los ganadores. (Esto puede ser aplicable en cualquier línea de pobreza). Por lo tanto, se puede flexibilizar el supuesto del “anonimato” o del “velo de ignorancia” del tradicional análisis de asistencia social, donde los resultados se juzgan únicamente por medio de los cambios en la distribución marginal (Carneiro et ál., 2001).

Cabría esperar heterogeneidad en los impactos de los programas de lucha contra la pobreza. Los criterios de elegibilidad imponen costos diferenciales en los participantes. Por ejemplo, los ya mencionados ingresos laborales obtenidos por los participantes en el programa de asistencia para desempleados, o los esquemas de transferencia de dinero condicional (a través de la pérdida de ganancias por el empleo de menores) serán diferentes en función de las competencias y las condiciones locales del mercado laboral. Para la economía política de las políticas de lucha contra la pobreza es muy importante obtener más información acerca de esta heterogeneidad, que además puede indicar la necesidad de políticas complementarias para proteger mejor a los perdedores.

Es fácil tener en cuenta la heterogeneidad de los impactos en términos de las características observables, agregando efectos de interacción con la variable de tratamiento ficticia (*dummy*), como en la ecuación (5.1). Pero sorprendentemente esto está alejado de la práctica universal. También se puede tener en cuenta la heterogeneidad latente, mediante un estimador de coeficientes aleatorios en el que la estimación de impacto (el coeficiente de la variable de tratamiento ficticia [*dummy*]) contiene un componente estocástico (es decir, $\mu_i^T \neq \mu_i^C$ en el término de error de la ecuación 4). Al aplicar este tipo de estimador a los datos de la evaluación de *PROGRESA*, Djebbari y Smith (2005) hallaron que pueden rechazar de manera

convinciente el supuesto de efectos comunes de las evaluaciones anteriores. Cuando existe esa heterogeneidad, generalmente se desean distinguir los impactos marginales de los impactos medios. De acuerdo con Björklund y Moffitt (1987), el efecto marginal del tratamiento se puede definir como la ganancia media de las unidades a las que les resulta indiferente participar o no. Para esto es necesario que se modele de manera explícita el problema de elección que enfrentan los participantes (Björklund y Moffitt, 1987; Heckman y Navarro-Lozano, 2004). También es posible que se desee estimar la distribución conjunta de Y^T e Y^C , y se puede encontrar un método para ello en Heckman et ál. (1997a).

Sin embargo, es cuestionable la relevancia que puedan tener en el contexto actual los modelos de elección que surgen de esta literatura. Los modelos provienen principalmente de la literatura sobre la evaluación de programas de capacitación y de otros programas en países desarrollados, donde la selección se considera como una elección individual, entre las personas elegibles. Este enfoque no se ajusta fácilmente a lo que se sabe acerca de muchos programas de lucha contra la pobreza en los países en desarrollo, donde las elecciones de los políticos y los administradores parecen ser más importantes que las elecciones de las personas elegibles para participar.

Esto habla de la necesidad de una caracterización teórica más fuerte del problema de selección en los trabajos futuros. Un ejemplo de un intento de esta índole es el modelo de Galasso y Ravallion (2005) para la asignación de un programa descentralizado de lucha contra la pobreza. El modelo que proponen se centra en el problema de elección pública que enfrenta el gobierno central y en el problema de acción colectiva local que enfrentan las comunidades, donde las elecciones de participación individual se consideran un problema insignificante.

Dichos modelos también indican variables instrumentales para identificar los impactos y analizar su heterogeneidad.

Cuando el problema de las políticas es la decisión de ampliar o reducir determinado programa en el margen, el estimador clásico del impacto medio sobre los tratados (mediante métodos experimentales o NX) resulta de poco interés. En la sección 7, se consideró el problema de intentar estimar el impacto marginal más allá de lo que dura la exposición del programa para los tratados, con el ejemplo de la comparación entre los que “permanecen” y los que “abandonan” el programa de asistencia para desempleados (Ravallion et ál., 2005). Se puede ver otro ejemplo en el estudio realizado por Behrman et ál. (2004) de los impactos sobre las destrezas cognitivas y condiciones sanitarias de niños, alcanzados por una mayor exposición a un programa preescolar en Bolivia. Los autores proporcionan una estimación del impacto marginal que supone una mayor duración del programa comparando los efectos acumulativos de las diferentes duraciones mediante un estimador de correspondencia. En tales casos, la selección dentro del programa no es un problema, y ni siquiera se necesitan datos sobre las unidades que nunca participaron. El método de diseños de discontinuidad analizado en la sección 6 (en su forma no paramétrica) y en la sección 8 (en su forma paramétrica de IV) también ofrece una estimación de la ganancia marginal de un programa; es decir, de la ganancia que se obtiene cuando se amplía (o reduce) mediante un pequeño cambio en el punto de corte de elegibilidad.

Comprender en profundidad los factores que determinan los resultados de las evaluaciones *ex post* también puede ayudar a simular *ex ante* los posibles impactos en el diseño de las políticas o del programa. Naturalmente, las simulaciones *ex ante* requieren muchos más

supuestos acerca de cómo funciona una economía.⁵⁸ En la medida de lo posible, sería bueno que esos supuestos estén basados en el conocimiento acumulado obtenido mediante rigurosas evaluaciones *ex post*. Por ejemplo, al combinar las opciones de un diseño de evaluación aleatorio con un modelo de educación estructural y al utilizar el diseño aleatorio para la identificación, se pueden ampliar en gran medida las preguntas relacionadas con las políticas acerca del diseño de *PROGRESA*, en comparación con las que puede responder una evaluación convencional (Todd y Wolpin, 2002; Attanasio et ál. 2004; de Janvry y Sadoulet, 2006). La literatura dedicada a este tema ha revelado que un cambio de presupuesto neutral del subsidio de matriculación de la escuela primaria a la secundaria podría haber generado beneficios netos en los niveles académicos obtenidos, al aumentar la proporción de niños que pasan a la escuela secundaria. Si bien *PROGRESA* tuvo un impacto en la escolaridad, éste podría haber sido mayor. No obstante, se debe tener en cuenta que este tipo de programa tiene dos objetivos: aumentar la escolaridad (para reducir la pobreza en el futuro) y reducir la pobreza actual, a través de transferencias con objetivos seleccionados. En la medida en que volver a centrar los subsidios en la escolaridad secundaria pueda reducir el impacto en la pobreza actual (al aumentar los ingresos que no se percibirán por el empleo de menores), debería analizarse con más detalle si está justificado el cambio en el diseño del programa.

10. Conclusiones

Este estudio nos deja dos lecciones a tener en cuenta en las futuras evaluaciones de los programas de lucha contra la pobreza. En primer lugar, ninguna herramienta de evaluación se puede considerar ideal en todas las circunstancias. Si bien la aleatorización puede ser una herramienta eficaz para evaluar el impacto, no es ni indispensable ni suficiente para lograr una buena evaluación. A veces los economistas no han cuestionado lo suficiente sus estrategias de identificación NX, aun así en la práctica se pueden encontrar medios creíbles para aislar al

⁵⁸ Se puede obtener una aproximación útil de los métodos *ex ante* en Bourguignon y Ferreira (2003). Todd y Wolpin (2006) dan varios ejemplos, incluso para un programa de subsidio escolar, con los datos obtenidos de *PROGRESA*.

menos una parte de la variación exógena en un programa ubicado de manera endógena. Se pueden obtener buenas evaluaciones prácticamente de todas las herramientas disponibles, en algunos casos combinando métodos: aleatorizando algunos aspectos y utilizando métodos econométricos para tratar los elementos no aleatorizados, utilizando elementos de un programa como fuente de variables instrumentales, o combinando métodos de pareo con observaciones longitudinales para intentar eliminar los errores de correspondencia que surgen de datos inexactos. Generalmente las buenas evaluaciones requieren, además, que el evaluador esté involucrado desde el inicio del programa y que esté bien informado de cómo funciona en la práctica. Las características del diseño y la implementación del programa pueden ofrecer importantes pistas para evaluar el impacto con medios NX.

En segundo lugar, aun si dejamos a un lado las cuestiones de validez interna, es poco probable que las herramientas del análisis contrafactual en variables de resultado bien definidas sean suficientes para brindar información para políticas y proyectos de desarrollo futuros. El contexto donde se implementa el programa puede influir en gran medida en los resultados. Esto indica la necesidad de comprender en profundidad *por qué* un programa tiene impacto o no lo tiene. Exige además un enfoque ecléctico basado en varias fuentes, que cuando sea posible incluya replicaciones en varios contextos y que pruebe los supuestos presentados en el diseño de un programa. Para lograrlo se puede, por ejemplo, realizar un seguimiento de las variables intermedias o bien utilizar teorías complementarias o evidencia externa a la evaluación. A los fines de obtener lecciones útiles para las políticas de lucha contra la pobreza, se necesita un conjunto de parámetros de impacto más vasto que el tradicionalmente implementado en la práctica de evaluación, incluida la posibilidad de distinguir los impactos en ganadores de los impactos en perdedores a cualquier nivel de vida determinado. En última instancia la elección de los parámetros a estimar en una evaluación debe depender de la pregunta sobre políticas que se pretende responder. Para las autoridades responsables de la formulación de políticas esto es algo habitual, pero demasiado a menudo parece no existir para los evaluadores.

Figura 1: Región de soporte común

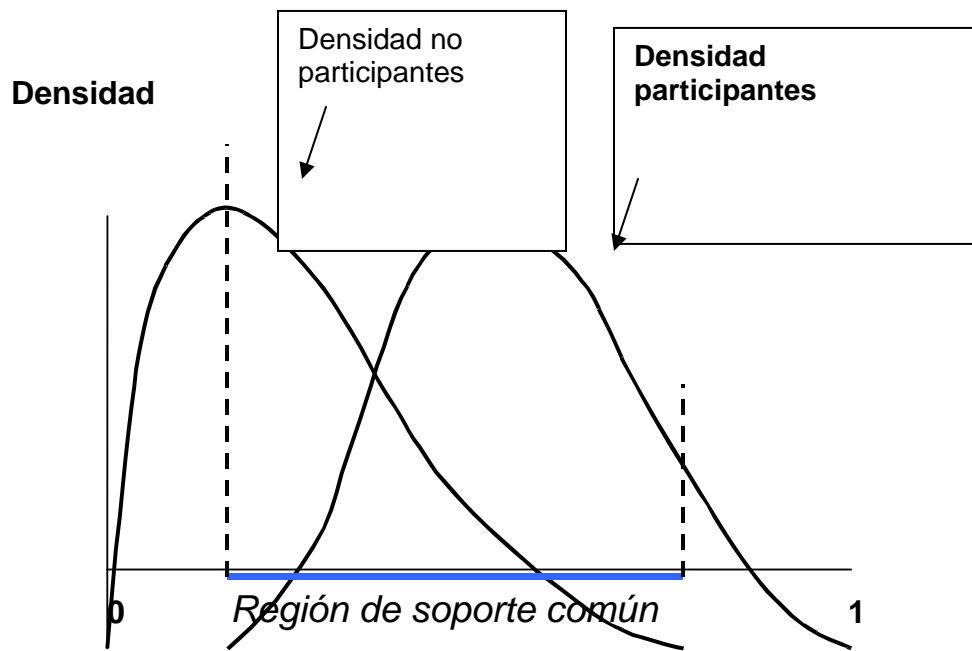
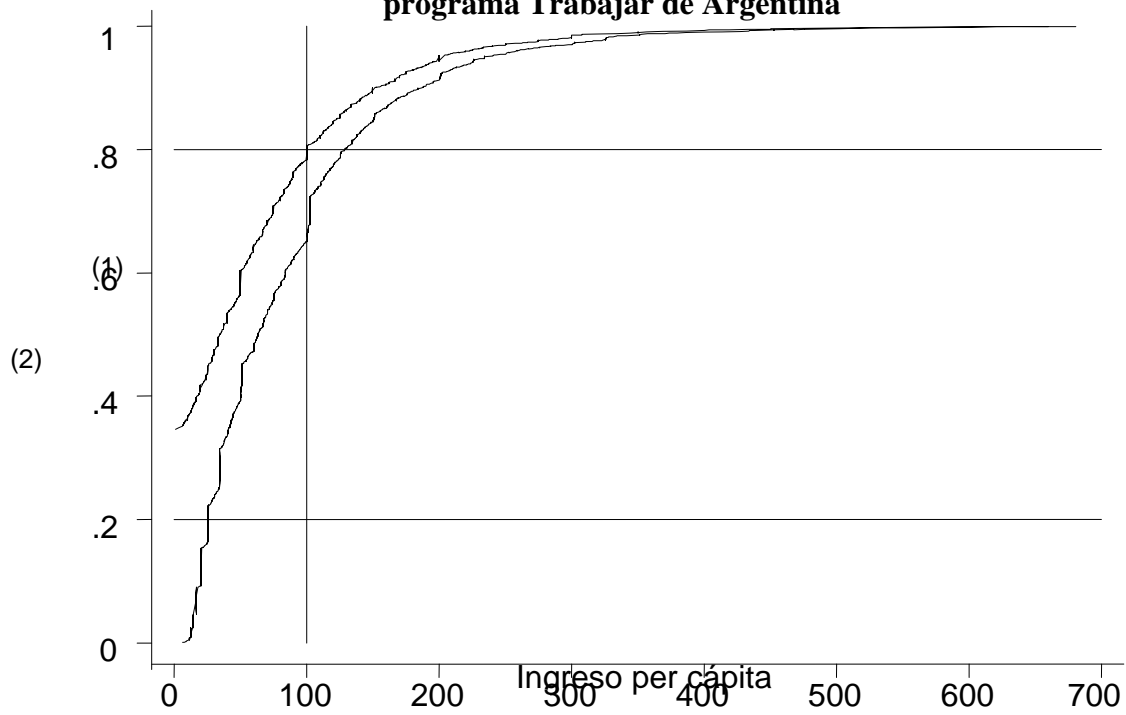


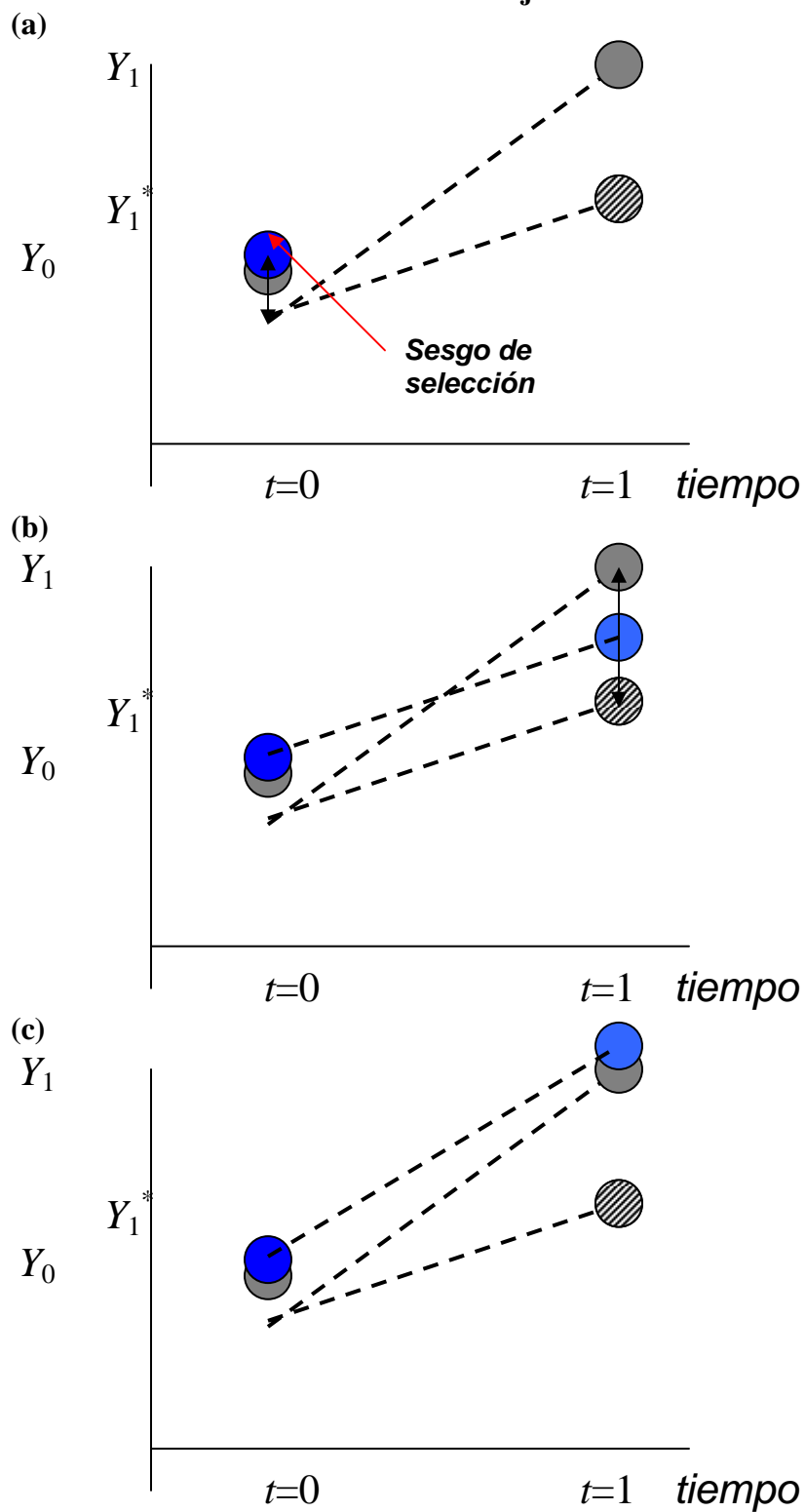
Figura 2: Impacto sobre la pobreza de los desembolsos de dinero pertenecientes al programa Trabajar de Argentina



- (1) Muestra de participantes antes de la intervención (estimada)
- (2) Muestra de participantes después de la intervención (observada)

Fuente: Jalan y Ravallion (2003b).

Figura 3: Sesgos en estimaciones de doble diferencia para un programa contra la pobreza con objetivos seleccionados.



Referencias

- Abadie, Alberto y Guido Imbens, 2006, "Large Sample Properties of matching Estimators for Average Treatment Effects," *Econometrica* 74(1): 235-267.
- Agodini, Roberto y Mark Dynarski, 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *Review of Economics and Statistics* 86(1): 180-194.
- Altonji, Joseph, Todd E. Elder y Christopher R. Taber, 2005a, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy* 113(1): 151-183.
- _____, _____ y _____, 2005b, "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schools," *Journal of Human Resources* 40(4): 791-821.
- Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King y Michael Kremer, 2002, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92(5): 1535-1558.
- Angrist, Joshua y Jinyong Hahn, 2004, "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1): 58-72.
- Angrist, Joshua, Guido Imbens y Donald Rubin, 1996, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, XCI: 444-455.
- Angrist, Joshua y Alan Krueger, 2001, "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15(4): 69-85.
- Angrist, Joshua y Victor Lavy, 1999, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2): 533-575.
- Ashenfelter, Orley, 1978, "Estimating the Effect of Training Programs on Earnings," *Review of Economic Studies* 60: 47-57.
- Atkinson, Anthony, 1987, "On the Measurement of Poverty," *Econometrica*, 55: 749-64.
- Attanasio, Orazio, Costas Meghir y Ana Santiago, 2004, "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," Working Paper EWP04/04, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.

- Attanasio, Orazio y A. Marcos Vera-Hernandez, 2004. "Medium and Long Run Effects of Nutrition and Child Care: Evaluation of a Community Nursery Programme in Rural Colombia," Working Paper EWP04/06, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.
- Basu, Kaushik, Ambar Narayan y Martin Ravallion, 2002, "Is Literacy Shared Within Households?" *Labor Economics* 8: 649-665.
- Battistin, Erich y Enrico Rettore, 2002, "Testing for Programme Effects in a Regression Discontinuity Design with Imperfect Compliance," *Journal of the Royal Statistical Society A*, 165(1): 39-57
- Behrman, Jere, Yingmei Cheng y Petra Todd, 2004, "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach," *Review of Economics and Statistics*, 86(1): 108-32.
- Behrman, Jere, Piyali Sengupta y Petra Todd, 2002, "Progressing through PROGESA: An Impact Assessment of a School Subsidy Experiment in Mexico," mimeo, University of Pennsylvania.
- Bertrand, Marianne, Esther Duflo y Sendhil Mullainathan, 2004, "How Much Should we Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119(1): 249-275.
- Besley, Timothy y Anne Case, 2000, "Unnatural Experiments? Estimating the Incidence of Endogeneous Policies," *Economic Journal* 110(November): F672-F694.
- Bhalla, Surjit, 2002, *Imagine There's No Country: Poverty, Inequality and Growth in the Era of Globalization*, Washington DC.: Institute for International Economics.
- Björklund, Anders y Robert Moffitt, 1987, The Estimation of Wage Gains and Welfare Gains in Self-Selection, *Review of Economics and Statistics* 69(1): 42-49.
- Bloom, Howard S., 1984, "Accounting for No-shows in Experimental Evaluation Designs," *Evaluation Review* 8: 225-246.
- Bourguignon, François y Francisco Ferreira, 2003, "Ex-ante Evaluation of Policy Reforms Using Behavioural Models," en Bourguignon, F. and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.
- Bourguignon, Francois, Anne-Sophie Robilliard y Sherman Robinson, 2003. "Representative

- Versus Real Households in the Macro-Economic Modeling of Inequality,” Working Paper 2003-05, DELTA, Paris.
- Buddelmeyer, Hielke y Emmanuel Shoufias, 2004, “An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA,” Policy Research Working Paper 3386, World Bank, Washington DC.
- Burtless, Gary, 1985, “Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment,” *Industrial and Labor Relations Review*, Vol. 39, pp. 105-115.
- _____, 1995, “The Case for Randomized Field Trials in Economic and Policy Research,” *Journal of Economic Perspectives* 9(2): 63-84.
- Carneiro, Pedro, Karsten Hansen y James Heckman, 2001, “Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies,” *Swedish Economic Policy Review* 8: 273-301.
- Carvalho, Soniya y Howard White, 2004, “Theory-Based Evaluation: The Case of Social Funds,” *American Journal of Evaluation* 25(2): 141-160.
- Case, Anne y Angus Deaton, 1998, “Large Cash Transfers to the Elderly in South Africa,” *Economic Journal* 108:1330-61.
- Chase, Robert, 2002, “Supporting Communities in Transition: The Impact of the Armenian Social Investment Fund,” *World Bank Economic Review*, 16(2): 219-240.
- Chen, Shaohua y Martin Ravallion, 2004, “Household Welfare Impacts of WTO Accession in China,” *World Bank Economic Review*, 18(1): 29-58.
- Chen, Shaohua, Ren Mu y Martin Ravallion, 2006, “Longer-Term Impacts of a Poor-Area Development Project,” Policy Research Working Paper, World Bank, Washington DC.
- Cook, Thomas, 2001. “Comments: Impact Evaluation: Concepts and Methods,” en O. Feinstein and R. Piccioto (eds), *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.
- Deaton, Angus, 1995, “Data and Econometric Tools for Development Analysis,” en Jere Behrman and T.N. Srinivasan (eds), *Handbook of Development Economics, Volume 3*, Amsterdam: North-Holland.
- _____, 1997, *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*, Baltimore: Johns Hopkins University Press for the World Bank.

- _____, 2005, "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)," *Review of Economics and Statistics*, 87(1): 1-19.
- Dehejia, Rajeev, 2005, "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics* 125(1-2), 355-364.
- Dehejia, Rajeev y S. Wahba, 1999, "Causal Effects in NX Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062.
- De Janvry, Alain y Elisabeth Sadoulet, 2006, "Making Conditional Cash Transfer Programs More Efficient: Designing for Maximum Effect of the Conditionality," *World Bank Economic Review* 20(1): 1-29.
- Diaz, Juan Jose y Sudhanshu Handa, 2004, "An Assessment of Propensity Score Matching as a NX Impact Estimator: Evidence from a Mexican Poverty Program," mimeo, University of North Carolina Chapel Hill.
- Djebbari, Habiba y Jeffrey Smith, 2005, "Heterogeneous Program Impacts of PROGRESA," mimeo, Laval University y University of Michigan.
- Dubin, Jeffrey A. y Douglas Rivers, 1993, "Experimental Estimates of the Impact of Wage Subsidies," *Journal of Econometrics*, 56(1/2): 219-242.
- Duflo, Esther, 2001, "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4): 795-813.
- _____, 2003, "Grandmothers and Granddaughters: Old Age Pension and Intrahousehold Allocation in South Africa," *World Bank Economic Review* 17(1): 1-26.
- Duflo, Esther y Michael Kremer, 2005, "Use of Randomization in the Evaluation of Development Effectiveness," en George Pitman, Osvaldo Feinstein and Gregory Ingram (eds.) *Evaluating Development Effectiveness*, New Brunswick, NJ: Transaction Publishers.
- Dubin, Jeffrey A. y Douglas Rivers, 1993, "Experimental Estimates of the Impact of Subsidies," *Journal of Econometrics*, 56(1/2), 219-242.
- Foster, James, J. Greer y Erik Thorbecke, 1984, "A Class of Decomposable Poverty Measures," *Econometrica*, 52: 761-765.
- Fraker, Thomas y Rebecca Maynard, 1987, "The Adequacy of Comparison Group Designs for

- Evaluations of Employment-Related Programs,” *Journal of Human Resources* 22(2): 194-227.
- Frankenberg, Elizabeth, Wayan Suriastini y Duncan Thomas, 2005, “Can Expanding Access to Basic Healthcare Improve Children’s Health Status? Lessons from Indonesia’s ‘Midwife in the Village’ Program,” *Population Studies* 59(1): 5-19.
- Frölich, Markus, 2004, “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators,” *Review of Economics and Statistics*, 86(1): 77-90.
- Gaiha, Raghav y Katushi Imai, 2002, “Rural Public Works and Poverty Alleviation: The Case of the Employment Guarantee Scheme in Maharashtra,” *International Review of Applied Economics* 16(2): 131-151.
- Galasso, Emanuela y Martin Ravallion, 2004, “Social Protection in a Crisis: Argentina’s *Plan Jefes y Jefas*,” *World Bank Economic Review*, 18(3): 367-399.
- _____ y _____, 2005, “Decentralized Targeting of an Anti-Poverty Program,” *Journal of Public Economics*, 85: 705-727.
- Galasso, Emanuela, Martin Ravallion y Agustin Salvia, 2004, “Assisting the Transition from Workfare to Work: Argentina’s Proempleo Experiment”, *Industrial and Labor Relations Review*, 57(5):.128-142.
- Galiani, Sebastian, Paul Gertler, y Ernesto Schargrotsky, 2005, “Water for Life: The Impact of the Privatization of Water Services on Child Mortality,” *Journal of Political Economy*, 113(1): 83-119.
- Gertler, Paul, 2004. “Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment” *American Economic Review, Papers and Proceedings* 94(2): 336-41.
- Glazerman, Steven, Dan Levy y David Myers, 2003, “NX versus Experimental Estimates of Earnings Impacts,” *Annals of the American Academy of Political and Social Sciences* 589: 63-93.
- Glewwe, Paul, Michael Kremer, Sylvie Moulin y Eric Zitzewitz, 2004, “Retrospective vs. Prospective Analysis of School Inputs: The Case of Flip Charts in Kenya,” *Journal of Development Economics* 74: 251-268.
- Godtland, Erin, Elizabeth Sadoulet, Alain De Janvry, Rinku Murgai y Oscar Ortiz, 2004, “The Impact of Farmer Field Schools on Knowledge and Productivity: A Study of Potato

- Farmers in the Peruvian Andes,” *Economic Development and Cultural Change*, 53(1): 63-92.
- Hahn, Jinyong, 1998, “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica* 66: 315-331.
- Hahn, Jinyong, Petra Todd y Wilbert Van der Klaauw, 2001, “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design,” *Econometrica* 69(1): 201-209.
- Hausman, Jerry, 1978, “Specification Tests in Econometrics,” *Econometrica* 46: 1251-1271.
- Heckman, James, 1979, “Sample Selection Bias as a Specification Error,” *Econometrica* 47(1): 153-161.
- Heckman, James y Joseph Hotz, 1989, “Choosing Among Alternative NX Methods for Estimating the Impact of Social Programs: The Case of Manpower Training,” *Journal of the American Statistical Association* 84: 862-874.
- Heckman, James, Hidehiko Ichimura y Petra Todd, 1997b, “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *Review of Economic Studies* 64(4), 605-654.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith y Petra Todd, 1998, “Characterizing Selection Bias using Experimental Data,” *Econometrica* 66, 1017-1099.
- Heckman, James, Robert Lalonde y James Smith, 1999, “The Economics and Econometrics of Active Labor Market Programs,” *Handbook of Labor Economics, Volume 3*, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.
- Heckman, James, L. Lochner y C. Taber, 1998, “General Equilibrium Treatment Effects,” *American Economic Review Papers and Proceedings* 88: 381-386.
- Heckman, James y Salvador, Navarro-Lozano, 2004, “Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models,” *Review of Economics and Statistics* 86(1): 30-57.
- Heckman, James y Richard Robb, 1985, “Alternative Methods of Evaluating the Impact of Interventions”, en J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press.
- Heckman, James y Jeffrey Smith, 1995, “Assessing the Case for Social Experiments,” *Journal of Economic Perspectives* 9(2): 85-110.

- Heckman, James, Jeffrey Smith y N. Clements, 1997a, "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for heterogeneity in Programme Impacts," *Review of Economic Studies* 64(4), 487-535.
- Hirano, Keisuke y Guido Imbens, 2004, "The Propensity Score with Continuous Treatments," *En Missing Data and Bayesian Methods in Practice*, Wiley forthcoming.
- Hirano, Keisuke, Guido Imbens y G. Ridder, 2003, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71: 1161-1189.
- Hoddinott, John y Emmanuel Skoufias, 2004, "The Impact of PROGRESA on Food Consumption," *Economic Development and Cultural Change* 53(1): 37-61.
- Holland, Paul, 1986, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81: 945-960.
- Holtz-Eakin, D., W. Newey y H. Rosen, 1988, "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56: 1371-1395.
- Imbens, Guido, 2000, "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 83: 706-710.
- _____, 2004, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics* 86(1): 4-29.
- Imbens, Guido y Joshua Angrist, 1994, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62(2): 467-475.
- Jacob, Brian y Lars Lefgren, 2004, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics* 86(1): 226-44
- Jacoby, Hanan G., 2002, "Is There an Intrahousehold 'Flypaper Effect'? Evidence from a School Feeding Programme," *Economic Journal* 112(476): 196-221.
- Jalan, Jyotsna y Martin Ravallion, 1998, "Are There Dynamic Gains from a Poor-Area Development Program?" *Journal of Public Economics*, 67(1), 65-86.
- _____ y _____, 2002, "Geographic Poverty Traps? A Micro Model of Consumption Growth in Rural China", *Journal of Applied Econometrics* 17(4): 329-346.
- _____ y _____, 2003a, "Does Piped Water Reduce Diarrhea for Children in Rural India?" *Journal of Econometrics* 112: 153-173.
- _____ y _____, 2003b, "Estimating Benefit Incidence for an Anti-poverty Program using Propensity Score Matching," *Journal of Business and Economic Statistics*,

21(1): 19-30.

- Kapoor, Anju Gupta, 2002, *Review of Impact Evaluation Methodologies Used by the Operations Evaluation Department over 25 Years*, Operations Evaluation Department, World Bank.
- Katz, Lawrence F., Jeffrey R. Kling y Jeffrey B. Liebman, 2001, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, 116(2): 607-654.
- Korinek, A., Mistiaen, J.A., Ravallion, M., 2006, "Survey Nonresponse and the Distribution of Income." *Journal of Economic Inequality*, 4(2): 33-55.
- Lalonde, Robert, 1986, "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76: 604-620.
- Lanjouw, Peter y Martin Ravallion, 1999, "Benefit Incidence and the Timing of Program Capture," *World Bank Economic Review*, 13(2): 257-274.
- Lee, Donghoon, 2005, "An Estimable Dynamic General Equilibrium Model of Work, Schooling, and Occupational Choice," *International Economic Review*, 46(1): 1-34.
- Lokshin, M. y M. Ravallion, 2000, "Welfare Impacts of Russia's 1998 Financial Crisis and the Response of the Public Safety Net." *Economics of Transition*, 8(2): 269-295.
- Manski, Charles, 1990, "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings* 80: 319-323.
- _____, 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60: 531-542.
- Miguel, Edward y Michael Kremer, 2004, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica* 72(1): 159-217.
- Moffitt, Robert, 1991, "Program Evaluation with NX Data," *Evaluation Review*, 15(3): 291-314.
- _____, 2001, "Policy Interventions, Low-Level Equilibria and Social Interactions," en Steven Durlauf and H. Peyton Young (eds) *Social Dynamics*, Cambridge Mass.: MIT Press.
- _____, 2003, "The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs," Cemmap Working Paper, CWP23/02, Department of Economics, University College London.
- Murgai, Rinku y Martin Ravallion, 2005, "Is a Guaranteed Living Wage a Good

- Anti-Poverty Policy?" Policy Research Working Paper, World Bank, Washington DC.
- Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, y Jose Luis Evia, 2002, "An Impact Evaluation of Education, Health y Water Supply Investments by the Bolivian Social Investment Fund," *World Bank Economic Review*, 16: 241-274.
- Paxson, Christina y Norbert R. Schady, 2002, "The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru," *World Bank Economic Review* 16: 297-319.
- Piehl, Anne, Suzanne Cooper, Anthony Braga y David Kennedy, 2003, "Testing for Structural Breaks in the Evaluation of Programs," *Review of Economics and Statistics* 85(3): 550-558.
- Pitt, Mark y Shahidur Khandker, 1998, "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy* 106: 958-998.
- Pitt, Mark, Mark Rosenzweig y Donna Gibbons, 1995, "The Determinants and Consequences of the Placement of Government Programs in Indonesia, en: D. van de Walle and K. Nead, eds., *Public spending and the poor: Theory and evidence* (Johns Hopkins University Press, Baltimore).
- Rao, Vijayendra y Ana Maria Ibanez, 2005, "The Social Impact of Social Funds in Jamaica: A Mixed Methods Analysis of Participation, Targeting and Collective Action in Community Driven Development," *Journal of Development Studies* 41(5): 788-838.
- Rao, Vijayendra y Michael Woolcock, 2003. "Integrating Qualitative and Quantitative Approaches in Program Evaluation," en F. Bourguignon and L. Pereira da Silva (eds.), *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.
- Ravallion, Martin, 1996, "Issues in Measuring and Modeling Poverty," *Economic Journal*, 106: 1328-44.
- _____, 2000, "Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved," *World Bank Economic Review* 14(2): 331-45.
- _____, 2003a, "Assessing the Poverty Impact of an Assigned Program," en Bourguignon, F. and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.

- _____, 2003b, "Measuring Aggregate Economic Welfare in Developing Countries: How Well do National Accounts and Surveys Agree?," *Review of Economics and Statistics*, 85: 645-652.
- _____, 2004a, "Who is Protected from Budget Cuts?" *Journal of Policy Reform*, 7(2): 109-22.
- _____, 2004b, "Looking beyond Averages in the Trade and Poverty Debate," Policy Research Working Paper 3461, World Bank, Washington DC.
- _____, 2005, "Poverty Lines," in *New Palgrave Dictionary of Economics*, 2nd edition, Larry Blume and Steven Durlauf (eds) London: Palgrave Macmillan.
- Ravallion, Martin y Shaohua Chen, 2005, "Hidden Impact: Household Saving in Response to a Poor-Area Development Project," *Journal of Public Economics*, 89: 2183-2204.
- Ravallion, Martin y Gaurav Datt, 1995. "Is Targeting through a Work Requirement Efficient? Some Evidence for Rural India," en D. van de Walle y K. Nead (eds) *Public Spending and the Poor: Theory and Evidence*, Baltimore: Johns Hopkins University Press.
- Ravallion, Martin, Emanuela Galasso, Teodoro Lazo y Ernesto Philipp, 2005, "What Can Ex-Participants Reveal About a Program's Impact?" *Journal of Human Resources*, 40(Winter): 208-230.
- Ravallion, Martin, Dominique van de Walle y Madhur Gaurtam, 1995, "Testing a Social Safety Net," *Journal of Public Economics*, 57(2): 175-199.
- Ravallion, Martin y Quentin Wodon, 2000, "Does Child Labor Displace Schooling? Evidence on Behavioral Responses to an Enrolment Subsidy," *Economic Journal* 110: C158-C176.
- Rosenbaum, Paul y Donald Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.
- Rosenzweig, Mark y Kenenth Wolpin, 1986, "Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions," *American Economic Review* 76, 470-82.
- Rubin, Donald B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Education Psychology* 66: 688-701.
- _____, 1979, "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74: 318-328.

- Rubin, Donald B., y N. Thomas, 2000, "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association* 95, 573-585.
- Sadoulet, Elizabeth, Alain de Janvry y Benjamin Davis, 2001, "Cash Transfer Programs with Income Multipliers: PROCAMPO in Mexico," *World Development* 29(6): 1043-56.
- Sala-i-Martin, Xavier, 2002, "The World Distribution of Income (Estimated from Individual Country Distributions)," NBER Working Paper No. W8933.
- Schultz, T. Paul, 2004, "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program," *Journal of Development Economics*, 74(1): 199-250.
- Skoufias, Emmanuel, 2005, *PROGRESA and Its Impact on the Welfare of Rural Households in Mexico*, Research Report 139, International Food Research Institute, Washington DC.
- Smith, Jeffrey y Petra Todd, 2001, "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review*, 91(2), 112-118.
- _____, y _____, 2005a, "Does Matching Overcome LaLonde's Critique of NX Estimators?" *Journal of Econometrics*, 125(1-2): 305-353.
- _____, y _____, 2005b, "Rejoinder," *Journal of Econometrics*, 125(1-2): 365-375.
- Thomas, Duncan, Elizabeth Frankenberg, Jed Friedman *et ál.*, 2003, "Iron Deficiency and the Well-Being of Older Adults: Early Results from a Randomized Nutrition Intervention," Paper Presented at the Population Association of America Annual Meetings, Minneapolis.
- Todd, Petra, 2006, "Evaluating Social programs with Endogeneous Program Placement and Selection of the Treated," *Handbook of Development Economics Volume 4*, edited by Robert E. Evenson y T. Paul Schultz, Amsterdam, North-Holland.
- Todd, Petra y Kenneth Wolpin, 2002, "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and fertility: Assessing the Impact of a School Subsidy Program in Mexico," Penn Institute for Economic Research Working Paper 03-022, Department of Economics, University of Pennsylvania.
- _____, y _____, 2006, "Ex-Ante Evaluation of Social Programs," mimeo, Department of Economics, University of Pennsylvania.

- van de Walle, Dominique, 2002, "Choosing Rural Road Investments to Help Reduce Poverty," *World Development* 30(4).
- _____, 2004, "Testing Vietnam's Safety Net," *Journal of Comparative Economics*, 32(4): 661-679.
- van de Walle, Dominique y Dorothy-Jean Cratty, 2005. "Do Aid Donors Get What they Want? Microevidence on Fungibility," Policy Research Working Paper 3542, World Bank.
- Vella, Francis y Marno Verbeek, 1999, "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business and Economic Statistics* 17(4): 473-478.
- Watts, H.W., 1968, "An Economic Definition of Poverty," in D.P. Moynihan (ed.), *On Understanding Poverty*. New York, Basic Books.
- Weiss, Carol, 2001, "Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs," en O. Feinstein and R. Piccioto (eds), *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.
- Woodbury, Stephen y Robert Spiegelman, 1987, "Bonuses to Workers and Employers to Reduce Unemployment," *American Economic Review*, 77, 513-530.
- Wooldridge, Jeffrey, 2002, *Econometric Analysis of Cross-Section and Panel Data*, Cambridge, Mass.: MIT Press.