El Enfoque Experimental de Economía del Desarrollo

The Experimental Approach to Development Economics

Abhijit V. Banerjee y Esther Duflo[1]

Noviembre de 2008

Documento de trabajo NBER 14467

http://www.nber.org/papers/w14467

Traducción: Enrique A. Bour

Resumen

Los experimentos aleatorios se han convertido en una herramienta popular en la investigación de economía del desarrollo y han sido objeto de varias críticas. En este documento se examina la literatura reciente y se analizan los puntos fuertes y las limitaciones de este enfoque en teoría y en la práctica. Sostenemos que la principal virtud de los experimentos aleatorios es que, en virtud de la estrecha colaboración entre investigadores y ejecutores, permiten la estimación de parámetros que de otro modo no sería posible evaluar. Analizamos las preocupaciones que se han planteado en relación con los experimentos y, en general, llegamos a la conclusión de que, si bien son reales, a menudo no son específicos de los experimentos. Concluimos discutiendo la relación entre teoría y experimentos.

Abhijit V. Banerjee
MIT
Department of Economics
Cambridge, MA
y NBER

Esther Duflo
Department of Economics
MIT, Cambridge,
y NBER

[1] Se agradece a Guido Imbens por muchas conversaciones útiles.

En los últimos años se ha producido una verdadera explosión de experimentos aleatorios en la economía del desarrollo y con ello, quizás inevitablemente, una oleada creciente de críticas. Casi todas las críticas son bienintencionadas, reconociendo los beneficios de tales experimentos y sugiriendo al mismo tiempo que no hay que olvidar que existen muchas cuestiones importantes que los experimentos aleatorios no pueden responder. Muchas de ellas tampoco son nuevas. De hecho, la mayoría de las objeciones estándar (y algunas no tan estándar) pueden hallarse en una sola pieza seminal de James Heckman, escrita hace más de una década (Heckman, 1992).

Gran parte de esta crítica ha sido útil, incluso cuando no estamos del todo de acuerdo con ella, tanto para ayudarnos a pensar en las fortalezas y limitaciones de lo que se ha hecho, como para aclarar hacia dónde debe dirigirse el campo a continuación. Sin embargo, argumentaremos que gran parte de esta crítica pasa por alto (o al menos no hace suficiente hincapié en) las principales razones por las que ha habido tanto entusiasmo en torno a la investigación experimental en economía del desarrollo. A continuación volveremos a las diversas críticas, en parte para aclararlas y calificarlas, y en parte para argumentar que, debido a un reconocimiento imperfecto de lo apasionante del programa experimental, existe la tendencia a establecer falsas oposiciones entre la labor experimental y otras formas de investigación.

1. La promesa de los experimentos

La investigación experimental en economía del desarrollo, al igual que las investigaciones anteriores en economía laboral y economía de la salud, partió de la preocupación por la identificación fiable de los efectos de los programas frente a canales de causalidad complejos y múltiples. Los experimentos permiten variar un factor a la vez y por lo tanto proporcionan estimaciones "internamente" válidas del efecto causal. La labor experimental realizada a mediados del decenio de 1990 (por ejemplo, Glewwe, Kremer y Moulin, 2007; Glewwe, Kremer, Moulin y Zitzewitz,2004; Banerjee, Jacob y Kremer, 2005), tenía por objeto responder a preguntas muy básicas sobre la función de producción educativa: ¿tiene importancia un mejor acceso a los insumos (libros de texto, rotafolios en las clases, una menor proporción de alumnos por maestro) para los resultados escolares (asistencia, resultados de exámenes) y, de ser así, en qué medida?

Esta investigación produjo una serie de resultados sorprendentes. La mejora del acceso a libros de texto de uno por cada cuatro o más estudiantes a uno por cada dos no afecta la puntuación media de los exámenes (Glewwe, Kremer y Moulin, de próxima aparición); tampoco lo hace la reducción a la mitad de la relación profesor-alumno (Banerjee, Jacob y Kremer, 2005). Por otra parte, también se pueden obtener resultados sorprendentemente positivos: Un estudio sobre el tratamiento de los parásitos intestinales en las escuelas de Kenya (Miguel y Kremer, 2004), demostró que un tratamiento antiparasitario que cuesta 49 centavos por niño por año puede reducir el ausentismo en una cuarta parte. Esto se debe en parte a factores externos: los parásitos se transmiten al caminar descalzos en lugares donde otros niños

infectados por parásitos han defecado. En consecuencia, en términos de aumento de la asistencia, la eliminación de parásitos es casi veinte veces más eficaz que la contratación de un maestro adicional (el costo de un año de educación extra para el niño era de 3,25 dólares con la eliminación de parásitos, en comparación con unos 60 dólares para el programa de maestros adicionales, a pesar de que al maestro adicional se le pagaba sólo unos 25 dólares al mes), aunque ambos "funcionan" en el sentido de generar mejoras estadísticamente significativas.

Lo que esta investigación estaba dejando claro es que a nivel de la eficacia de los ingredientes individuales de la función de producción educativa, nuestra intuición (o la teoría económica per se) era poco probable que nos ayudara mucho, ¿cómo podríamos saber, a priori, que la desparasitación es mucho más eficaz que la contratación de un profesor? En términos más generales, un boletín del laboratorio de Acción contra la Pobreza Abdul Latif Jameel compara el costo por año de educación extra inducida a través de una serie de estrategias diferentes (J-PAL, 2005). Los costos varían ampliamente, entre 3,50 dólares por año-niño extra para desparasitación y 6.000 dólares por año-niño extra para el componente de educación primaria de PROGRESA, el Programa de Transferencia Condicionada de Dinero de México. Algunos de estos programas (como los de PROGRESA) pueden tener también otros objetivos. Pero para aquellos cuyo objetivo principal es aumentar la educación, está claro que algunos son mucho más baratos que otros. Incluso excluyendo PROGRESA, el costo por año extra de educación inducida oscila entre $3.25 y más de $200. Por lo tanto, incluso al comparar entre programas para lograr el mismo objetivo, las tasas de retorno de la inversión pública están lejos de ser igualadas.

Además, quedó claro que los economistas no eran los únicos que no tenían ni idea: las organizaciones encargadas de la ejecución no estaban mucho mejor informadas. Por ejemplo, la ONG que financiaba la intervención antiparasitaria también se mostró inicialmente entusiasmada con la idea de dar a los niños uniformes escolares, aunque una evaluación aleatoria demostró que el costo de un año extra de escolarización de los niños que daban a los niños un uniforme gratuito resultó ser de 100 dólares por año de escolarización.

De esta experiencia surgieron varias conclusiones importantes. En primer lugar, la formulación de políticas eficaces requiere hacer juicios sobre la eficacia de los componentes individuales de los programas, sin mucha orientación de conocimiento a priori. En segundo lugar, sin embargo, también es difícil aprender sobre estos componentes individuales a partir de datos de observación (es decir, no experimentales). La razón es que los datos observacionales sobre la función de producción educativa a menudo provienen de sistemas escolares que han adoptado un "modelo" determinado, que consiste en más de un insumo. Por lo tanto, la variación de los insumos escolares que observamos proviene de intentos de cambiar el modelo, lo cual, por muy buenas razones, implica hacer múltiples cambios al mismo tiempo. Un buen ejemplo es la "Operación Pizarra" en la India (Chin, 2005), un programa de modernización de las escuelas que implica la contratación de nuevos maestros y la entrega

de material de enseñanza y aprendizaje a las escuelas en forma simultánea. Todos los programas de educación posteriores en la India (el Programa de Educación Primaria de Distrito, el Sarva Shiksha Abhiyan) han tenido esta característica. Si bien hay excepciones importantes (por ejemplo, el hecho de que el tamaño de las clases cambie de manera discontinua con la matriculación en Israel permite una evaluación limpia del impacto de sólo el tamaño de las clases, véase Angrist y Lavy (1992)), esto significa que muchos de los conocimientos pertinentes para las políticas que requieren la observación de los efectos de la variación de componentes individuales de un conjunto pueden no estar disponibles en los datos de observación. Esta es una primera motivación para los experimentos.

Una implicancia inmediata de esta observación es que, dado el costo fijo de organizar un experimento y el hecho de que los experimentos requieran necesariamente algún tiempo cuando hay que retrasar la ejecución del programa (para poder utilizar los resultados), vale la pena hacer múltiples experimentos al mismo tiempo en la misma población, que evalúan variantes potenciales alternativas del programa. Por ejemplo, el Banco Mundial proporcionó dinero a comités escolares para que contrataran a maestros adicionales con contratos breves a fin de reducir el tamaño de las clases en el primer grado en Kenia. Cuando trabajaron con el sistema escolar para establecer una evaluación del programa, los investigadores no se limitaron a asignar todo el programa a escuelas de tratamiento seleccionadas al azar (Duflo, Dupas y Kremer, 2008). Se introdujeron dos dimensiones adicionales de variación: la capacitación del comité escolar que recibía el dinero para vigilar al maestro adicional, y el seguimiento por logros previos. Así pues, este diseño permite estimar el impacto de la reducción del tamaño de las clases sin cambios en la pedagogía, el mérito relativo de los maestros jóvenes y adicionales con contratos breves en comparación con maestros regulares y experimentados, funcionarios públicos, el papel que pueden desempeñar los comités escolares debidamente habilitados y el impacto del seguimiento por rendimiento en la escuela primaria. Al igual que en Banerjee, Jacob y Kremer (2005), aunque en un contexto muy diferente, el estudio no encuentra que la reducción del tamaño de las clases sin ningún otro cambio tenga un impacto significativo. Sin embargo, mostró un fuerte impacto positivo del cambio de maestro regular a maestro contratado, un impacto positivo y significativo de la reducción del tamaño de la clase cuando se combina con el fortalecimiento del comité escolar y, para un tamaño de clase determinado, un fuerte beneficio del seguimiento de los estudiantes, tanto para los más débiles como para los más sólidos.

Otros "experimentos de tratamiento múltiple" son Banerjee, Cole, Duflo y Linden (2007), (educación de recuperación y aprendizaje asistido por computadora), Duflo, Dupas, Kremer y Sinei (2006) y Dupas (2007), (diversas estrategias de prevención del VIH-SIDA entre los adolescentes), Banerjee, Barnerji, Duflo, Glennerster y Khemani (2008) (experimentos de información y movilización en escuelas primarias de la India), Banerjee, Duflo, Glennerster y Kothari (2008) (factores de demanda y oferta para mejorar las tasas de inmunización en la India), Gine, Karlan y Zinman

(2008) (dos estrategias para ayudar a los fumadores a dejar de fumar), y muchos otros.

Una observación relacionada es que desde el punto de vista de la construcción de una base de conocimientos utilizable, es necesario un proceso de aprendizaje dinámico: En primer lugar porque los resultados experimentales son a menudo sorprendentes y, por lo tanto, requieren una mayor aclaración. Duflo, Kremer y Robinson (2008a, b) reflejan exactamente ese proceso iterativo, en el que se llevó a cabo una sucesión de experimentos sobre el uso de fertilizantes durante un período de varios años, y cada resultado dio lugar a la necesidad de probar una serie de nuevas variaciones para comprender mejor los resultados de la anterior.

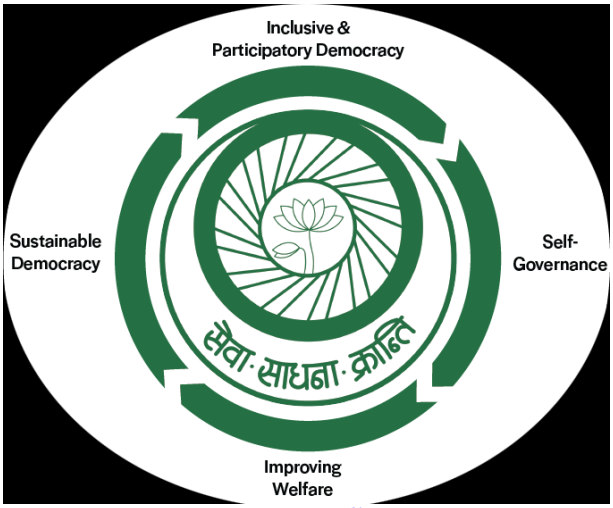
En segundo lugar, desde el punto de vista del aprendizaje óptimo, a menudo vale la pena probar una intervención amplia primero para ver si hay un efecto general y luego, si se encuentra que funciona, profundizar en sus componentes individuales, como una forma de entender qué parte del programa amplio funciona.[2] Los experimentos de políticas a menudo se detienen en el primer paso: un ejemplo es el popular programa Progresa-Oportunidades-Prospera, en México, que combinaba una transferencia de efectivo a las familias pobres condicionada al "buen comportamiento" (inversiones en educación y salud preventiva), con transferencias a las mujeres, y alguna mejora de las instalaciones educativas y sanitarias. El programa se ha reproducido en muchos países, a menudo junto con una evaluación aleatoria. Pero sólo en un estudio en curso en Marruecos se forman y comparan diferentes grupos de tratamiento, a fin de evaluar la importancia de las tan elogiadas condicionalidades. En este experimento, un grupo de aldeas recibe una transferencia puramente incondicional, un grupo recibe una transferencia de "condicionalidad débil", en la que los requisitos de asistencia sólo son verificados por los maestros, y dos grupos reciben una variante más estricta de la condicionalidad (en un grupo, la asistencia de los niños es supervisada por inspectores; en el otro, se verifica diariamente con un dispositivo de reconocimiento de huellas dactilares).

Si bien todo esto parece obvio en retrospectiva, fue sólo después de las primeras experiencias que tanto investigadores como las organizaciones de ejecución con las que trabajaron apreciaron plenamente lo que todo esto significaba para ellos. Desde el punto de vista de las organizaciones quedó claro que era valioso establecer relaciones relativamente a largo plazo con los investigadores, de modo que la experimentación pudiera constituir un proceso de aprendizaje continuo y se pudieran diseñar múltiples experimentos de interés mutuo. En otras palabras, se hizo menos hincapié en evaluaciones puntuales, en las que se trae al investigador para evaluar un programa específico que la organización ya ha decidido evaluar. Esta es una diferencia con la literatura de evaluación de los Estados Unidos o el Canadá, donde, con algunas excepciones importantes (por ejemplo, Angrist, Lang y Oreopoulos,

[2] O lo contrario: pasar de una intervención a la vez al paquete completo tiene sentido cuando sus premisas son que alguna combinación funcionará, mientras que lo contrario es mejor cuando es generalmente escéptico.

2009), los programas que se van a evaluar son elegidos principalmente por los organismos de ejecución, y los investigadores son evaluadores.

Desde el punto de vista de los investigadores, esto ofrecía la posibilidad de pasar del papel de evaluador al de co-experimentador, con un papel importante en la definición de lo que se evalúa. En otras palabras, al investigador se le ofrecía ahora la opción de definir la pregunta que debía responder, basándose en su conocimiento de lo que se sabía y en la teoría recibida. Por ejemplo, Seva Mandir, una ONG de Rajastán (India) con la que Banerjee y Duflo mantenían una relación de larga data, estaba interesada en mejorar la calidad de sus escuelas informales. Su idea inicial era poner en práctica un programa de incentivos para los maestros basado en resultados de exámenes. Sin embargo, los resultados de Glewwe, Ilias y Kremer (2003) los convencieron de que un peligro de los incentivos para los maestros sería enseñar para el examen u otra manipulación a corto plazo de las puntuaciones de los exámenes. Entonces decidieron poner en marcha un programa de incentivos basado en la presencia de los profesores. Para medir la asistencia, en una zona muy poco poblada donde las escuelas son de difícil acceso, Duflo y Hanna propusieron utilizar cámaras con indicación de fecha y hora. Aunque al principio Seva Mandir se sorprendió un poco por la sugerencia, aceptaron probarla. En las escuelas del programa (las "escuelas de cámara"), los maestros se tomaban una foto a sí mismos y a sus estudiantes dos veces al día (por la mañana y por la tarde), y su salario se calculaba como una función (no lineal) del número de días que asistían. Los resultados, de los que se informó en Duflo, Hanna y Ryan (2007), fueron bastante sorprendentes: la ausencia de los maestros se redujo de 40 a 20 puntos porcentuales, y el rendimiento de los estudiantes también mejoró. Seva Mandir quedó convencida de estos resultados y decidió continuar el programa. Sin embargo, no abandonaron la esperanza de mejorar la motivación intrínseca de los profesores. En lugar de extender el programa de cámaras en todas sus escuelas de inmediato, decidieron continuar con él en las escuelas donde ya habían sido introducidos, y pasar algún tiempo experimentando con otros programas, tanto en escuelas con cámaras como en escuelas sin ellas. Con Sendhil Mullainathan, realizaron una lluvia de ideas sobre las formas de motivar a los maestros. Una idea era dar a cada niño un diario para escribir todos los días basado en el trabajo hecho en la escuela. En los días en que el estudiante o el profesor estuviera ausente, el diario debía permanecer en blanco o ser tachado. Se suponía que los padres debían mirar el diario cada semana. La esperanza era que registraran la cantidad de ausencias de maestros y niños que había. Esto, resultó no tener éxito: los padres aparentemente partían de una opinión tan



Seva Mandir

baja de la escuela que el diario tendía a persuadirlos de que algo estaba pasando - los padres tienen una opinión más alta de las escuelas con diario que de las que no lo tienen, y no había ningún impacto en la presencia de profesores. Sin embargo, los diarios eran muy populares tanto entre los estudiantes como entre los profesores, e indujeron a los profesores a trabajar más duro cuando estaban presentes. Los resultados de las pruebas mejoraron en las escuelas de diarios. Por lo tanto, parece que los diarios fracasaron como herramienta para mejorar la presencia de los maestros, pero tuvieron éxito como herramienta pedagógica. Sin embargo, como esta no fue una hipótesis planteada en el diseño experimental inicial, puede ser sólo un accidente estadístico. Así pues, aunque Seva Mandir pondrá ahora cámaras en todas las escuelas (después de varios años, siguen teniendo un gran impacto en la presencia y en las puntuaciones de las pruebas), llevará a cabo un nuevo experimento con los diarios para ver si los resultados sobre la pedagogía persisten.

Una consecuencia importante de este proceso ha sido la creciente toma de conciencia en la comunidad de investigadores de que el elemento más importante del enfoque experimental puede residir en la facultad, cuando se trabaja con un asociado de ejecución amistoso, de variar los elementos individuales del tratamiento de manera que nos ayude a responder a preguntas conceptuales (aunque pertinentes para la política) que nunca podrían ser respondidas de manera fiable de otra manera.[3] Un ejemplo elocuente es Berry (2008). Aunque los incentivos basados en la participación y el rendimiento escolar se han hecho muy populares, no está claro si los incentivos deben dirigirse a los niños (como en los programas evaluados en Angrist, Lang y Oreopoulos (2008) y Angrist y Lavy (2002)) o a los padres (como en Kremer, Miguel y Thornton (2007)). Si la familia fuera totalmente eficiente la elección del objetivo no debería marcar una diferencia, pero de lo contrario podría hacerlo. Para responder a esta pregunta, Berry trabajó con Pratham en los barrios bajos de Delhi para diseñar un programa en el que se proporcionaban incentivos a los estudiantes (o a sus padres) (en forma de juguetes o dinero) basados en la mejora de la lectura del niño. Encontró que para los estudiantes inicialmente débiles, recompensar al niño es más efectivo en términos de mejorar los resultados de los exámenes que recompensar a los padres. Evidentemente, sin poder variar quién recibe los incentivos dentro del mismo contexto y en el mismo experimento, este estudio no habría sido posible.

Así pues, los experimentos están emergiendo como una poderosa herramienta para testear teorías en manos de aquellos con suficiente creatividad. Karlan y Zinman (2005) es un ejemplo. El proyecto se llevó a cabo en colaboración con un prestamista sudafricano que otorga pequeños préstamos a prestatarios de riesgo a elevadas tasas de interés. El experimento se diseñó para ensayar los pesos relativos de la carga de reembolso ex post (incluido el riesgo moral) y la selección adversa ex ante en el incumplimiento de los préstamos. A los posibles prestatarios con el mismo riesgo

[3] Si bien la limitación de trabajar con una organización de ejecución limita el conjunto de preguntas que se pueden hacer, en relación con lo que se puede hacer en un experimento de laboratorio, el realismo adicional del entorno parece ser una enorme ventaja.

observable se les ofrece al azar una tasa de interés alta o baja en una carta inicial. Los individuos deciden entonces si piden prestado a la tasa de "oferta" de la licitación. De los que solicitan una tasa más alta, a la mitad se les ofrece al azar una nueva tasa de interés "de contrato" más baja cuando realmente otorgan el préstamo, mientras que la mitad restante continúa a la tasa de la oferta. Las personas no sabían de antemano que la tasa de contrato podía diferir de la tasa de oferta. Los investigadores compararon entonces el rendimiento de la devolución de préstamos en los tres grupos. La comparación de los que respondían a la alta tasa de interés de oferta con los que respondían a la baja tasa de interés de oferta en la población que recibía la misma baja tasa de contrato permite identificar el efecto de selección adversa, mientras que la comparación de los que se enfrentaban a la misma tasa de oferta pero con diferentes tasas de contrato identifica el efecto de carga de la devolución.

El estudio encontró que las mujeres exhiben selección adversa pero los hombres exhiben riesgo moral. El hecho de que esta diferencia haya sido inesperada plantea un problema para el estudio (¿es una casualidad estadística o un fenómeno real?) pero su contribución metodológica es indiscutible. La idea básica de variar los precios ex post y ex ante para identificar los diferentes parámetros se ha reproducido desde entonces en varios estudios diferentes. Ashraf, Berry y Shapiro (2007) y Cohen y Dupas (2007) la explotan para comprender la relación entre el precio pagado por un bien de protección de la salud y su utilización. El aumento del precio podría afectar la utilización a través de un efecto de selección (los que compran a un precio más alto se preocupan más) o un "efecto de costo hundido psicológico". Para separar estos efectos, se aleatorizan el precio de oferta así como el precio real pagado. El efecto de que el precio de la oferta mantenga el precio real fijo identifica el efecto de selección, mientras que la variación del precio real (con un precio de oferta fijo) reduce el efecto de los costos hundidos. Ashraf y otros (2007) estudian esto para un producto de purificación de agua, mientras que Cohen y Dupas (2007) se dedican a los mosquiteros. En ninguno de los dos estudios hay muchas pruebas de un efecto psicológico de costo hundido. La variación experimental fue clave aquí, y no sólo para evitar el sesgo: en el mundo es poco probable que observemos un gran número de personas que se enfrenten a diferentes precios de oferta pero al mismo precio real. Este tipo de experimentos recuerdan la motivación de los primeros experimentos sociales (como los experimentos de impuesto sobre la renta negativo), que tenían por objeto obtener distintas variaciones de salarios e ingresos para estimar los efectos de renta y sustitución, que no estaban disponibles en los datos de observación (Heckman, 1992).

Otros ejemplos de este tipo de trabajo son los experimentos destinados a evaluar si existe una demanda de productos de compromiso: estos productos podrían ser demandados por personas con problemas de autocontrol. Ashraf, Karlan y Yin (2006) trabajaron con una institución de microfinanciación de Filipinas para ofrecer a sus clientes un producto de ahorro que les permitiera optar por comprometerse a no retirar el dinero antes de que se alcanzara un objetivo específico de tiempo o cantidad. Gine, Karlan y Zinman (2008) trabajaron con la misma organización para

invitar a los fumadores que quisieran dejar de fumar a que se comprometan con un "contrato": el dinero de una cuenta de ahorros especial se perdería si no superan un análisis de orina de fumadores después de varias semanas. En ambos casos, estos fueron diseñados por economistas para resolver un problema del mundo real, pero con una fuerte motivación teórica. El hecho de que se trataba de nuevas ideas que provenían de los investigadores hizo que fuera natural establecer una evaluación aleatoria: ya que eran de naturaleza experimental, los participantes normalmente estaban felices de probarlas primero con un subconjunto de sus clientes/beneficiarios.

Estos dos grupos de ejemplos se concentran en la conducta individual. Los experimentos también pueden ser establecidos para entender la forma en que funcionan las instituciones. Un ejemplo es el de Bertrand, Djankov, Hanna y Mullainathan (2009), que establecieron un experimento para comprender la estructura de la corrupción en el proceso de obtención del permiso de conducir en Delhi. Reclutan a personas que aspiran a obtener un permiso de conducir y establecen tres grupos, uno que recibe una bonificación por obtener rápidamente un permiso de conducir, otro que recibe clases de conducción gratuitas y un grupo de control. Se dan cuenta de que los que están en el grupo de "bonificación" obtienen sus licencias más rápido, pero los que reciben las lecciones de conducción gratuitas no. También descubren que es más probable que los del grupo de bonificación paguen a un agente para obtener la licencia (quien, conjeturan, a su vez soborna a alguien). También descubren que la contratación de un agente está correlacionada con una menor probabilidad de haber hecho un examen de conducir antes de obtener la licencia de conducir y de poder conducir. Si bien no parecen descubrir que los integrantes del grupo de bonificación que obtienen licencias sean sistemáticamente menos propensos a saber conducir que los del grupo de control (que sería la prueba de fuego de que la corrupción sí da lugar a una asignación ineficiente de licencias de conducir), este experimento aporta pruebas sugestivas de que la corrupción en este caso hace algo más que "engrasar las ruedas" del sistema.

La comprensión de que los experimentos son una opción fácilmente disponible también ha estimulado una mayor creatividad en la medición. Si bien hay muchos experimentos que utilizan métodos estándar y muchos trabajos no experimentales que han invertido mucho en mediciones (por ejemplo, Olken (2007b) sobre sobornos en la medición, Manski (2004, y muchos otros trabajos), sobre la medición de expectativas, las muestras biológicas en la Encuesta sobre Vida Familiar Indonesia y las Encuestas de Salud y Jubilación, etc.), la ventaja que ofrecen los experimentos es una alta tasa de reclamo y un problema específico de medición. En muchos estudios experimentales, una gran fracción de los que se pretende que se vean afectados por el programa se ven realmente afectados. Esto significa que el número de unidades sobre las que hay que recopilar datos para evaluar el impacto del programa no tiene por qué ser muy grande, y que los datos se suelen recopilar especialmente para el propósito del experimento. Así pues, es posible realizar una medición elaborada y costosa de los resultados.

Por el contrario, la mayoría de los estudios de observación cuasiexperimental se basan en algún tipo de cambio en la política para la identificación. Esos cambios de política suelen abarcar grandes poblaciones, lo que requiere el uso de grandes conjuntos de datos, que a menudo no se reúnen con ese fin específico. Además, aunque sea posible hacer ex post un ejercicio sofisticado de recopilación de datos dirigido específicamente al programa, generalmente es imposible hacerlo para la situación previa al programa. Esto impide el uso de una estrategia de diferencia en diferencias para este tipo de resultados.

Un ejemplo del tipo de datos que se recogieron en un entorno experimental es Olken (2007a). El objetivo era determinar si las auditorías o la vigilancia comunitaria eran formas eficaces de poner freno a la corrupción en los proyectos de construcción descentralizados. Así pues, era necesario obtener una medida fiable de los niveles reales de corrupción. Olken se centró en las carreteras, e hizo que los ingenieros cavaran perforaciones en la carretera para medir el material realmente utilizado. Luego comparó esto con el nivel de material que se comunicó que se había utilizado. La diferencia es una medida de la cantidad de material robado, o nunca comprado pero facturado, y por lo tanto una medida objetiva de corrupción. Olken demostró entonces que esta medida de "insumos faltantes" se ve afectada por la amenaza de auditorías, pero no por el fomento de una mayor asistencia a las reuniones de la comunidad, salvo en algunas circunstancias.

Otro ejemplo de reunión de datos innovadora se encuentra en Beaman, Chattopadhyay, Duflo, Pande y Topalova (2008). En ese documento se evalúan los efectos de la representación política obligatoria de la mujer en los consejos de aldea sobre la actitud de los ciudadanos hacia las mujeres dirigentes. Se trata de un experimento natural aleatorio en el sentido de que las aldeas fueron seleccionadas al azar (por ley) para ser "reservadas para las mujeres": en las aldeas "reservadas", sólo las mujeres podían ser elegidas jefas de aldea. Para obtener una medida del "gusto" por las mujeres líderes que no se viera empañada por el deseo del encuestado de complacer al entrevistador, el documento implementa "pruebas de asociación implícita", desarrolladas por psicólogos (Banaji, 2001). Si bien esas pruebas son utilizadas con frecuencia por los psicólogos, y su uso también ha sido propugnado por los economistas (Bertrand, Chugh y Mullainathan, 2005), no se habían aplicado en un entorno de campo en un país en desarrollo, y casi no había estudios que investigaran si esas actitudes están "bien conectadas" o pueden verse afectadas por las características del entorno. En el estudio también se utilizó otra medida de sesgo implícito hacia la mujer, inspirada por los politólogos. Los encuestados escuchan un discurso, supuestamente pronunciado por un líder de la aldea, pronunciado por una voz masculina o femenina, y se les pide que den su opinión al respecto. Los encuestados son seleccionados al azar para recibir el discurso masculino o femenino. La diferencia en las calificaciones dadas por los que reciben los discursos masculinos frente a los femeninos es una medida de discriminación estadística. El documento compara esta medida de discriminación entre aldeas reservadas y no reservadas.

Estos son sólo dos ejemplos de una literatura rica y creativa. Muchos experimentos de campo incluyen pequeños experimentos de laboratorio (juegos de dictador, elecciones sobre loterías, experimentos de tasas de descuento y juegos de bien público, etc.). También hay innovaciones en este sentido: por ejemplo, en las investigaciones en curso, con el fin de medir el "capital social", Erica Field y Rohini Pande distribuyeron billetes de lotería a los encuestados y dieron a los sujetos la opción de compartirlos con los miembros de sus grupos.

2. Preocupaciones sobre los experimentos

Como mencionamos, las preocupaciones sobre los experimentos no son nuevas. Sin embargo, muchas de ellas se basan en la comparación de los métodos experimentales, implícita o explícitamente, con otros métodos para tratar de aprender sobre la misma cosa. El mensaje de la sección anterior es que la mayor ventaja de los experimentos puede ser que nos lleven a terrenos en los que no se dispone de enfoques de observación. En esos casos, las objeciones planteadas por los críticos de la literatura experimental se consideran mejor como advertencias contra una interpretación excesiva de los resultados experimentales. Sin embargo, también hay casos en que se dispone de enfoques experimentales y de observación en formas relativamente comparables, en los que se plantea además la cuestión de qué enfoque adoptar. Por otra parte, existe preocupación por lo que los experimentos están haciendo a la economía del desarrollo como campo. En el resto de esta sección se enumeran esas objeciones y luego se examinan una por una.

2.1 Dependencia del medio ambiente

La dependencia del medio ambiente es un elemento central de generalizabilidad. Se plantea la pregunta de si obtendríamos el mismo resultado si realizamos el mismo experimento en un entorno diferente o, más exactamente, si el programa que se está evaluando tendría el mismo efecto si se aplicara en otra parte (no en el contexto de un experimento).

En realidad se trata de dos preocupaciones separadas: La primera y más obvia es que nos preocupe el impacto de las diferencias del entorno experimental en la eficacia del programa. Una virtud de los experimentos es que nos permiten evaluar el efecto medio del programa para una población específica sin suponer que el efecto del programa es constante entre los individuos. Pero si el efecto no es constante entre los individuos, es probable que varíe sistemáticamente con las covariables. Por ejemplo, los uniformes escolares seguramente no tendrán el mismo impacto en Noruega (donde cada niño que necesita uno, sin duda, lo tiene) que en Kenia. La cuestión es dónde trazar la línea: ¿México es más parecido a Noruega o más parecido a Kenya? La misma cuestión también se plantea dentro de un país. Es evidente que los conocimientos a priori sólo pueden ayudarnos aquí hasta cierto punto: la simple economía sugiere que los uniformes sólo tendrán efecto en las poblaciones en las que el salario medio no sea demasiado alto en relación con el precio de los

uniformes, pero ¿cuánto es demasiado alto? Si nuestras teorías son lo suficientemente buenas para saber esto, o estamos dispuestos a asumir que lo son, entonces probablemente ya no necesitamos experimentos: la teoría puede entonces ser lo suficientemente buena para darnos una idea de quién tiende a conseguir un uniforme, y quién no, y podríamos utilizar esta restricción para estimar de manera convincente modelos estructurales del impacto de los uniformes escolares. En otras palabras, sin supuestos, los resultados de los experimentos no pueden generalizarse más allá de su contexto; pero con suficientes supuestos, los datos observacionales pueden ser suficientes. Para argumentar a favor de los experimentos, tenemos que estar en algún punto intermedio.

Un segundo problema proviene de la preocupación por efectos de implementadores. En particular, cuanto más pequeña sea la organización ejecutora, mayor será la preocupación de que el efecto estimado del tratamiento refleje las características únicas del ejecutor. Una preocupación conexa expresada por Heckman (1992) es que los sitios u organizaciones que aceptan formar parte de un experimento pueden ser diferentes de otras. Por ejemplo, señala que varios sitios se negaron a participar en los experimentos del JTPA (Job Training Partnership Act) porque se oponían a la randomización.

Este problema puede mitigarse parcialmente proporcionando información detallada sobre la aplicación en la descripción de la evaluación, haciendo hincapié en el lugar que ocupa el programa evaluado dentro del plan de acción general de la organización (qué tamaño tenía la pieza evaluada en relación con lo que hacen, cómo se seleccionó el equipo de ejecución, qué decidió la elección del lugar, etc.). Evidentemente, para que los resultados sean algo más que una "prueba de concepto" inicial, el programa debe provenir de un programa suficientemente bien definido y comprendido como para que su ejecución se delegue de forma rutinaria a un gran número de equipos de ejecución individuales más o menos autosuficientes.

Todo esto es sin embargo muy vago y altamente subjetivo (¿qué es lo suficientemente grande? ¿cuán autosuficiente?, etc.). Para abordar ambas preocupaciones sobre generalización, es necesario llevar a cabo estudios de réplica reales. Se deben realizar experimentos adicionales en diferentes lugares, con diferentes equipos. Si tenemos una teoría que nos dice dónde es probable que los efectos sean diferentes, centramos los experimentos adicionales allí. Si no es así, lo ideal sería elegir ubicaciones aleatorias dentro del dominio relevante.

De hecho, actualmente existen varios estudios de replicación, aunque, como señaló Heckman, los lugares donde se realizan los experimentos no suelen elegirse al azar. El programa de enseñanza suplementaria ("*balsakhi*") evaluado por Banerjee y otros (2007), en realidad se llevó a cabo deliberadamente de manera simultánea en dos lugares distintos (Mumbai y Vadodara) trabajando con dos equipos de ejecución separados (ambos de la red de Pratham, pero bajo una gestión totalmente distinta). Los resultados resultaron ser en general coherentes. Análogamente, Bobonis,

Miguel y Sharma (2006) obtienen un impacto similar de una combinación de tratamientos antiparasitarios y suplementos de hierro en la asistencia escolar en el norte de la India que Miguel y Kremer (2004) encontraron en Kenya, y Bleakley (2007) encuentra resultados similares utilizando datos naturales del sur de los Estados Unidos a principios del siglo XX mediante un enfoque de experimentación natural.

El programa PROGRESA/Oportunidades fue replicado bajo diferentes nombres y con ligeras variantes en muchos países, y en algunos de ellos fue acompañado de una evaluación aleatoria (Colombia, Nicaragua, Ecuador y Honduras; Marruecos está en marcha). Los resultados fueron muy coherentes en todos los países.

Otros resultados no son reproducibles: Una campaña de información que movilizó a los comités de padres sobre cuestiones relacionadas con la educación y los alentó a utilizar un programa público que permite a los comités escolares contratar a maestros locales cuando las escuelas están superpobladas, tuvo un efecto positivo en los resultados del aprendizaje en Kenia pero no en la India (Banerjee, Banerji, Duflo, Glennerster y Khamani y otros, 2008; Duflo, Dupas Kremer, 2008). Y una intervención similar que buscaba dinamizar los comités de gestión de las unidades de salud en Uganda informó de un impacto masivo en resultados difíciles de afectar como la mortalidad infantil (Bjorkman y Svensson, 2007).

Además de la mera replicación, se generan conocimientos acumulativos a partir de experimentos conexos en diferentes contextos. El examen analítico de Kremer y Holla (2008) de 16 experimentos aleatorios sobre elasticidad-precio en salud y educación es un buen ejemplo. Volveremos sobre estos resultados con más detalle más adelante, pero el punto clave aquí es que estos experimentos cubren una amplia gama de bienes y servicios de educación y salud, en varios países. Un hilo común muy fuerte es la elasticidad extremadamente alta de la demanda de estos bienes con relación a su precio, especialmente en torno a cero (tanto en la dirección positiva como en la negativa). Si bien no son estrictamente réplicas de cada uno de ellos, esto muestra claramente el valor de los conocimientos acumulados en el aprendizaje de un fenómeno.

Sin embargo, está claro que se necesita mucha más investigación sobre replicación. A algunos les preocupa que haya pocos incentivos en el sistema para llevar a cabo estudios de replicación (ya que es posible que las revistas no estén tan dispuestas a publicar el quinto experimento sobre un tema determinado como el primero), y es posible que los organismos de financiación tampoco estén dispuestos a financiarlos. El uso extensivo de experimentos en economía es todavía reciente, por lo que no sabemos cuán grande puede ser el problema, aunque dadas las numerosas estimaciones publicadas sobre los rendimientos de la educación, por ejemplo, no somos demasiado pesimistas. La buena noticia es que se están llevando a cabo varios esfuerzos de replicación sistemática. Por ejemplo, un programa de transferencia de activos y capacitación dirigido a los muy pobres, diseñado originalmente por la ONG bangladesí BRAC, (descrito en detalle más adelante) se está evaluando actualmente

en Honduras, el Perú, Karnataka, Bengala Occidental, Bangladesh y el Pakistán. Cada país cuenta con un equipo de investigación diferente y un asociado local diferente para la ejecución. Actualmente se están realizando estudios sobre la sensibilidad de las tasas de interés que replican a Karlan y Zinman (2008) en Ghana, el Perú (en dos lugares distintos con dos asociados diferentes), México y Filipinas (en tres lugares distintos con dos asociados diferentes). Se están realizando evaluaciones del impacto del microcrédito simultáneamente en Marruecos, la India urbana, Filipinas (en tres lugares distintos) y México. Se está evaluando la capacitación empresarial en el Perú, la República Dominicana, la India urbana y México. Se están evaluando programas similares para fomentar el ahorro en el Perú, Filipinas, Ghana y Uganda. Parece, pues, que hay suficiente interés entre los organismos de financiación para financiar estos experimentos, y suficientes investigadores dispuestos a llevarlos a cabo. Por ejemplo, en el caso de los diversos experimentos de alta pobreza en curso, la Fundación Ford los está financiando todos, en un intento explícito de lograr una mayor comprensión del programa mediante su evaluación en varios lugares separados. Innovations for Poverty Action (una ONG fundada por Dean Karlan), que ha encabezado el esfuerzo para muchas de estas réplicas, está recibiendo la subvención, pero los equipos de investigación y los asociados en la ejecución son diferentes en cada país. Los diferentes equipos de investigación comparten estrategias e instrumentos de evaluación, para asegurarse de que los diferentes resultados representen diferencias en contextos, más que en estrategias de evaluación.

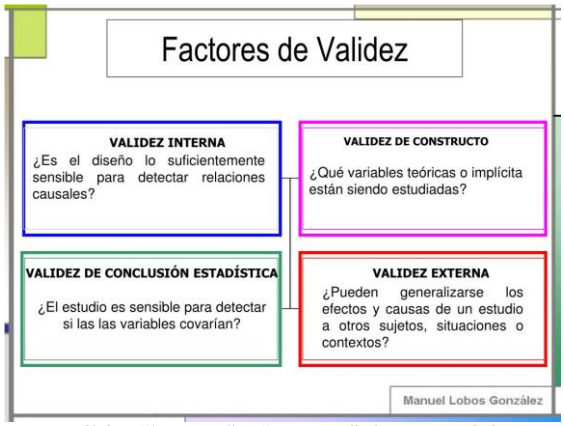
Esos estudios todavía están en curso, y sus resultados nos dirán mucho más sobre las condiciones en que los resultados de los programas dependen del contexto. Se necesitarán pruebas sistemáticas para saber si los resultados difieren entre los sitios. Un enfoque será tratar los diferentes sitios como covariables, y utilizar la prueba no paramétrica propuesta por Crump et al (2008) para comprobar si el efecto es diferente en cualquiera de los sitios. Si se detecta heterogeneidad, una prueba más potente sería si la heterogeneidad sigue existiendo después de tener en cuenta la heterogeneidad de las covariables. Otra forma de proceder sería ejecutar las regresiones no paramétricas propuestas por Crump et al (2008) y comprobar si el efecto del tratamiento condicionado por las covariables es igual para todos las dummies de los sitios. Si bien no es una propuesta directa de Crump et al (2008), sería una extensión directa. La cuestión, obviamente, no es que todos los resultados de la investigación experimental se generalicen, sino que tengamos una forma de saber cuáles sí y cuáles no. Si estuviéramos preparados para realizar suficientes experimentos en lugares suficientemente variados, podríamos aprender todo lo que quisiéramos saber sobre la distribución de los efectos del tratamiento en sitios condicionados a un determinado conjunto de covariables.

En cambio, no se puede hacer ninguna afirmación comparable sobre los estudios observacionales. Si bien tal vez sea posible identificar un cuasiexperimento particular que ofrezca de manera convincente el efecto de tratamiento correcto, parece muy improbable que ese cuasiexperimento pueda reproducirse en tantos entornos diferentes como se desee. Además, en los estudios observacionales es necesario suponer

que no hay confusión (es decir, que las hipótesis de identificación son válidas) en todos los estudios para poder compararlos. Si varios estudios observacionales dan resultados diferentes, una posible explicación es que uno o varios de ellos estén sesgados (este es el principio que subyace a una prueba de sobreidentificación), y otra es que los efectos del tratamiento sean efectivamente diferentes.

El problema de Identificación es un tema tratado por la Econometría

Sin embargo, a menudo se afirma -véase Rodrik (2008), por ejemplo- que la dependencia del medio ambiente es menos problemática en los estudios observacionales porque esos estudios abarcan zonas mucho más amplias y, por consiguiente, el efecto del "tratamiento" es un promedio en un gran número de entornos y, por lo tanto, más generalizable.[4] En este sentido, se sugiere que existe un equilibrio entre los estudios aleatorios más "internamente" válidos y los estudios de observación más "externamente" válidos.



Validez "interna" y "externa" de un modelo
Véase *Medición, muestreo y relaciones causales 32-37*

Sin embargo, esto no es necesariamente cierto. Una parte del problema se reduce a lo que significa ser generalizable: significa que si se toma la misma acción en un lugar diferente se obtendría el mismo resultado. ¿Pero qué acción y qué resultado? En los estudios de sección transversal que comparan, por ejemplo, diferentes tipos de inversiones, el hecho de que la acción fuera la misma y que los resultados se midieran de la misma manera debe tomarse basándose en la fe, una decisión de confiar en el juicio de aquellos que construyeron el conjunto de datos y reunieron una serie de programas bajo un encabezamiento general. Por ejemplo, "inversión en educación" podría significar varias cosas diferentes. Por lo tanto, la conclusión generalizable del estudio es, en el mejor de los casos, el impacto del promedio del conjunto de cosas que resultaron ser puestas en común al construir los datos agregados.

[4] Nótese que no todos los experimentos aleatorios son de pequeña escala. Por ejemplo, los programas de representación obligatoria que mencionamos anteriormente se implementaron en todo el país en la India. Mientras que Chattopadhyay y Duflo (2004) originalmente sólo examinaron dos estados (muy diferentes), Topalova y Duflo (2004) ampliaron el análisis a todos los principales estados de la India.

También hay un tema más sutil sobre las generalizaciones, que surge incluso cuando evaluamos programas individuales muy bien definidos. El hecho de que la evaluación de un programa utilice datos de una gran área, no significa necesariamente que la estimación del efecto del programa que obtenemos de esa evaluación sea un promedio de los efectos del programa en todos los diferentes tipos de personas que viven en esa gran área (o en todas las personas que son plausibles participantes en el programa). La forma en que estimamos el efecto del programa en esos casos es tratar primero de controlar cualquier diferencia observable entre los que están cubiertos por el programa y los que no lo están (por ejemplo, utilizando algún tipo de cotejo) y luego observar cómo se desempeñan los que están en el programa en relación con los que no lo están.

Pero es posible que una vez que emparejemos a los iguales, o bien casi todos los que están en un grupo emparejado en particular son participantes del programa o todos son no participantes. Existen varios métodos para tratar esta falta de coincidencia entre la distribución de participantes y no participantes (Heckman, Ichimura y Todd, 1997; Heckman, Ichimura, Smith y Todd, 1998; Rubin 2006-véase una reseña en Imbens y Wooldridge, 2008), pero en todos los casos, la estimación será impulsada enteramente por los subgrupos de la población en los que, incluso después del emparejamiento, haya suficientes participantes y no participantes, y estos subgrupos podrían ser enteramente no representativos. Y aunque podemos identificar las características observables de la población que impulsan la estimación del efecto del tratamiento (aunque esto rara vez se hace) no tenemos forma de saber cómo se comparan con el resto de la población en términos de inobservables. En palabras de Imbens y Wooldridge, "una característica potencial de todos estos métodos [que mejoran la superposición entre participantes y no participantes] es que cambian lo que se está estimando (...) Esto da lugar a una reducción de la validez externa, pero es probable que mejore la validez interna". Así pues, el equilibrio entre validez interna y externa también está presente en los estudios de observación. Por el contrario, mientras las tasas de cumplimiento entre los elegidos para el tratamiento en un experimento sean altas, sabemos que la población afectada es por lo menos representativa de la población elegida para el experimento. Como es bien sabido (véase Imbens y Angrist, 1994), el mismo punto se aplica también a las estimaciones de variable instrumental: los "cumplidores" en una estrategia de VI, para los que se identifica el efecto del programa, pueden ser un subconjunto pequeño y no representativo de la población de interés.

El punto señalado por Heckman (1992) aún se mantiene. Si las evaluaciones aleatorias sólo pueden llevarse a cabo en lugares muy específicos o con socios específicos, precisamente porque son aleatorias y no todos los socios están de acuerdo con la aleatoriedad, la repetición en muchos sitios no elimina este problema. Se trata de una objeción grave (estrechamente relacionada con el problema del cumplimiento que examinamos a continuación: es el cumplimiento a nivel de la organización) y difícil de refutar, ya que ninguna cantidad de datos podría asegurarnos completamente que no se trata de un problema. Nuestra experiencia es que, en el contexto de

los países en desarrollo, esto es cada vez menos un problema a medida que las evaluaciones aleatorias adquieren una mayor aceptación: se han completado proyectos de evaluación con organizaciones no gubernamentales internacionales, gobiernos locales y una serie de organizaciones no gubernamentales locales. Esto sólo mejorará si la evaluación aleatoria llega a ser recomendada por la mayoría de los donantes, ya que significará que la voluntad de cumplir con la randomización ya no distingue entre organizaciones.

Una cuestión más grave en nuestra experiencia es el hecho conexo de que lo que distingue a los posibles asociados para evaluaciones aleatorias es la competencia y la voluntad de ejecutar los proyectos según lo previsto. Estos pueden perderse cuando el proyecto se amplíe. Es importante reconocer este límite cuando se interpretan los resultados de las evaluaciones: descubrir que un programa determinado, cuando se implementa en algún lugar, tiene un efecto medio determinado deja abierto el problema de cómo ampliarlo. Hasta ahora no se han hecho suficientes esfuerzos para intentar la evaluación a "mediana escala" de programas que han tenido éxito a pequeña escala, donde estos problemas de implementación se harían evidentes.

Dicho esto, este problema tampoco está del todo ausente en los estudios observacionales, especialmente en los países en desarrollo. No todos los programas pueden ser evaluados de manera convincente con un estudio comparativo. A menudo se requieren grandes conjuntos de datos (especialmente si se quiere mejorar la validez externa centrándose en un área grande). En algunos casos, los datos se recopilan a propósito para la evaluación, a menudo con la ayuda de la oficina de estadística del país. En este caso, el país debe aceptar la evaluación de un programa grande. Los programas grandes son políticamente más sensibles a la evaluación que los programas piloto, ya que suelen tener una buena publicidad, por lo que los países pueden ser estratégicos con respecto a la elección de los programas a evaluar. En otros casos, se pueden utilizar encuestas periódicas a gran escala (como la encuesta NSS en la India, la encuesta Susenas en Indonesia, etc.). Pero no todos los países en desarrollo las tienen (aunque los conjuntos de datos como los del *Department of Homeland Security* (algo equivalente a los Ministerios del Interior, que están disponibles para muchos países, ciertamente han mejorado la cuestión). Por lo tanto, también existe un sesgo potencial (aunque muy diferente del de la evaluación aleatoria) en los tipos de países y programas que pueden ser evaluados con datos observacionales.

No se trata de que la generalizabilidad no sea un problema para el enfoque experimental/cuasiexperimental, pero obviamente no es menos importante para cualquier otro enfoque.

2.2 Cuestiones de cumplimiento

Ya hemos señalado que una alta tasa de cumplimiento facilita la interpretación del efecto del tratamiento y la generalización de los resultados. Los experimentos en

economía del desarrollo a menudo se han llevado a cabo mediante la asignación al azar de un conjunto de lugares o grupos (aldeas, barrios, escuelas) en los que la organización encargada de la ejecución tiene una confianza relativa en la capacidad de ejecución. Por consiguiente, a nivel de ubicación la tasa de aceptación es relativamente alta, a menudo del 100%. Cabe destacar que esto sólo significa que es probable que la muestra tratada sea un subconjunto aleatorio del conjunto de lugares que se seleccionaron para el programa. Por supuesto, no está garantizado que los individuos reales que se beneficiaron del tratamiento sean un subconjunto aleatorio de la población de esos lugares, pero se supone que la selección a este nivel refleja la selección que induciría un programa real (es decir, no sólo en condiciones experimentales), y que el tratamiento en el parámetro tratado de una estimación de VI utilizando la aldea "de tratamiento" como instrumento es el parámetro de interés.

Heckman (1992) se ocupó específicamente de la interpretación de experimentos aleatorios en los Estados Unidos en los que se ofrecía a los individuos la opción de participar en un programa de capacitación laboral. La aceptación fue baja y potencialmente muy seleccionada, lo que estaría bien si quisiéramos saber el efecto de ofrecer a la gente tal opción, pero no si el plan era hacerlo obligatorio, por ejemplo, para todos los beneficiarios de asistencia social. Cuestiones similares también surgen en algunos de los experimentos de países en desarrollo. Por ejemplo, el estudio de Karlan y Zinman (2007) sobre el efecto del acceso al crédito al consumo parte de una población de personas cuya solicitud de préstamo fue rechazada por el banco. Luego piden a los oficiales de préstamos que identifiquen una clase de rechazos marginales de esta población y "no rechazados" aleatoriamente de un grupo de ellos. Sin embargo, los oficiales de préstamos todavía tenían discreción y la usaron para rechazar a cerca de la mitad de los que no fueron rechazados. El experimento identifica el efecto de este crédito extra en la población de los que permanecieron "no rechazados": Parece haber aumentado la probabilidad de que la persona siga teniendo empleo, así como su ingreso. Sin embargo, si bien proporciona pruebas (muy valiosas) de que el crédito de consumo podría ser bueno para algunas personas, dada la naturaleza inusual de la población tratada (la doblemente no rechazada) existe cierta incertidumbre sobre qué hacer con la magnitud real del efecto.

Otro punto señalado por Heckman es que las evaluaciones aleatorias no son el mejor método para estudiar quiénes toman los programas una vez que se les ofrecen y por qué. Esto no es necesariamente el caso, ya que la aleatorización puede ser usada precisamente para aprender sobre temas de selección: Como ya se ha dicho, existen actualmente varios estudios en los que la aleatorización está diseñada específicamente para medir el efecto de selección, lo que sería muy difícil de hacer de otra manera (Karlan y Zinman, 2005; Ashraf, Berry y Shapiro, 2007; Cohen y Dupas, 2007), como ya se ha dicho. Para saber más sobre selección, Cohen y Dupas (2007) recogieron el nivel de hemoglobina de mujeres que compraron mosquiteros a diferentes precios. Se interesaron en saber si las mujeres que compran mosquiteros sólo cuando son gratuitos tienen menos probabilidades de padecer anemia. En otros estudios, aunque la evaluación no está diseñada específicamente para captar el efecto

de selección, la aceptación entre las personas a las que se les ofrece el programa es de especial interés, y los datos de referencia se recopilan específicamente para este efecto. Por ejemplo, en Ashraf, Karlan y Yin (2006), un resultado importante de interés es quién toma un dispositivo de autocontrol que ayuda a las personas a ahorrar.

En otros casos, la aceptación no es un problema, porque el tratamiento es un puro regalo, a diferencia de la oferta de capacitación, que no tiene valor a menos que alguien esté dispuesto a dedicarle tiempo. Por ejemplo, De Mel, McKenzie y Woodruff (2008) estudian el efecto de ofrecer a cada empresa de su muestra en Sri Lanka unos 200 dólares en forma de capital adicional. Encuentran un gran impacto en los ingresos de la empresa, equivalente a un rendimiento del capital del 5 al 7%. Cull, McKenzie y Woodruff (2007) repiten el mismo experimento en México y encuentran rendimientos aún mayores (20-35%) En ambos casos, el hecho de que las empresas destinatarias fueran muy pequeñas fue crucial: esto fue lo que hizo que casi todos estuvieran interesados en participar en el programa (aunque fuera un regalo, siempre hay algún costo de participación), y permitió que un regalo tan pequeño (que es todo lo que podían permitirse) tuviera un impacto discernible.

Sin embargo, a veces incluso un regalo puede ser rechazado, como descubrieron, para su sorpresa, Banerjee, Chattopadhyay, Duflo y Shapiro, que trabajan con el Bandhan Bank Ltd. para evaluar sus programas de ayuda a los muy pobres (una de las varias evaluaciones de este programa que mencionamos anteriormente). En el marco de este programa, se identifica a los aldeanos que son demasiado pobres para incorporarse a la red de microfinanciación mediante evaluaciones participativas de los recursos y otras investigaciones de seguimiento y luego se les ofrece un activo (por lo general un par de vacas, unas cuantas cabras o algún otro activo productivo) por un valor de entre 25 y 100 dólares sin ningún tipo de condicionante legal (aunque se espera que se encarguen de ello y de hacer un seguimiento), así como una asignación semanal y algo de capacitación. El objetivo es ver si el acceso al activo crea una mejora a largo plazo en su nivel de vida (o simplemente venden el activo y gastan rápidamente). El diseño de la evaluación asumió que todos los que se les ofrece el activo lo tomarán, lo que resultó no ser el caso. Una fracción significativa de los clientes (18%) rechazó la oferta: Algunos sospechaban, porque pensaban que era parte de un intento de convertirlos al cristianismo; otros pensaban que era un truco para hacerlos caer en una trampa de endeudamiento, que finalmente se les exigiría que lo pagaran; otros no dudaban de los motivos de Bandhan, pero no se sentían capaces de hacer un buen trabajo cuidando el activo y no querían sentirse avergonzados en el pueblo si lo perdían.

2.3 Cuestiones de aleatoriedad

El ejemplo de Bandhan refuerza un punto también planteado en Heckman (1992): que el hecho de que haya un experimento en curso podría generar efectos de selección que no se producirían en entornos no experimentales. Este es un ejemplo de los

efectos de Hawthorne de John Henry: el hecho de formar parte de un experimento (y de ser supervisado) influye en sus participantes. El hecho de que estos habitantes no estuvieran acostumbrados a que una organización privada anduviera por ahí regalando bienes fue claramente una parte de la razón por la que se produjo el problema. Por otra parte, es posible que Bandhan no hiciera el tipo de esfuerzo de relaciones públicas para informar a los aldeanos sobre el motivo por el que se estaba haciendo, precisamente porque no tenían previsto atender a toda la población de los muy pobres de cada aldea.

Efecto Hawthorne (véase Medición, muestreo y relaciones causales, pág. 34*)*

La mayoría de los experimentos, sin embargo, son cuidadosos para evitar este problema. La aleatoriedad que tiene lugar a nivel de ubicación puede ir a caballito de una expansión de la participación de la organización en estas áreas limitadas por el presupuesto y la capacidad administrativa, que es precisamente la razón por la que aceptan la aleatoriedad. Los limitados presupuestos gubernamentales y las diversas acciones de muchas ONG pequeñas hacen que las aldeas o las escuelas de la mayoría de los países en desarrollo estén acostumbradas a que algunas zonas reciban algunos programas y otras no, y cuando una ONG sólo presta servicios a algunas aldeas, lo consideran parte de la estrategia general de la organización. Cuando las zonas de control dan la explicación de que el programa sólo tenía presupuesto suficiente para un cierto número de escuelas, suelen estar de acuerdo en que la lotería era una forma justa de asignarlo; a menudo están acostumbrados a esa arbitrariedad y por eso la asignación al azar parece transparente y legítima.

Un problema que plantea el reconocimiento explícito de la aleatoriedad como una forma justa de asignar el programa es que los encargados de su aplicación encontrarán que la forma más fácil de presentarlo a la comunidad es decir que se prevé una ampliación del programa para las zonas de control en el futuro (sobre todo cuando es así, como en el diseño por etapas). Esto puede causar problemas si la anticipación del tratamiento lleva a los individuos a cambiar su comportamiento. Esta crítica se hizo en el caso de los programas PROGRESA, donde las aldeas de control sabían que eventualmente serían cubiertas por el programa.

Cuando es necesario para la evaluación que las personas no tengan conocimiento de que están excluidas del programa por el bien de la evaluación, los comités de ética suelen conceder una exención de la divulgación completa hasta que se complete la encuesta final, al menos cuando el hecho de ser estudiado en el grupo de control no presente ningún riesgo para el sujeto. En estos casos, a los participantes en el nivel de base no se les dice que hay una aleatoriedad real en juego. Esto ocurre con mayor frecuencia cuando la aleatorización tiene lugar a nivel individual (aunque algunas aleatorizaciones a nivel individual se llevan a cabo por medio de una lotería pública).

En este caso, simplemente se revela a los beneficiarios seleccionados que, por ejemplo, obtuvieron un préstamo que habían solicitado (Karlan y Zinman, 2007), o que el banco había decidido que el tipo de interés podía ser más bajo (Karlan y Zinman, 2005).

2.4 Efectos de equilibrio

Una cuestión conexa es lo que suele denominarse, de forma ligeramente confusa, efectos de equilibrio general (y preferimos llamarlo efectos de equilibrio, ya que el equilibrio general es esencialmente un concepto multimercado). Los efectos del programa que se encuentran en un pequeño estudio pueden no generalizarse cuando el programa se amplía a escala nacional (Heckman, Lochner y Taber 1999). Considérese, por ejemplo, lo que sucedería si se intenta ampliar un programa que muestra, en una aplicación experimental a pequeña escala, que las niñas económicamente desfavorecidas que reciben vouchers para ir a escuelas privadas terminan con una mejor educación y mayores ingresos. Cuando ampliamos el programa a nivel nacional, surgen dos desafíos: uno es que habrá hacinamiento en las escuelas privadas (y potencialmente un colapso de las escuelas públicas) y el otro es que los beneficios de la educación disminuirán debido al aumento de la oferta. Por ambas razones, las pruebas experimentales podrían exagerar los beneficios del programa de vales.

El voucher educativo de Milton Friedman

Este fenómeno de efectos de equilibrio plantea un problema que no tiene una solución perfecta. Sin embargo, es evidente que hay muchos casos en los que no esperamos enfrentarlo: si queremos saber qué estrategia para promover la inmunización es más rentable (un parto fiable o un parto seguro, más un pequeño incentivo para que la madre se acuerde de inmunizar a su hijo en el plazo previsto) para aumentar las tasas de inmunización y en qué medida (como en Banerjee, Duflo, Glennerster y Kothari (2008), por ejemplo) el método experimental no plantea ningún problema. El hecho de que la inmunización de todo el distrito no requiera muchas más enfermeras adicionales nos ayuda aquí porque podemos asumir que el precio de las enfermeras no subiría mucho, si es que lo hace. Por otra parte, si bien es útil saber que a quienes recibieron vouchers en Colombia les va mejor en términos de resultados educativos y de vida (véase Angrist, Bettinger, Bloom y Kremer, 2002; Angrist, Bettinger y Kremer, 2006), es difícil no preocuparse por el hecho de que un aumento de la oferta general de aptitudes que se produzca por la ampliación del programa de vouchers hará bajar el precio de las aptitudes. Después de todo, esa es precisamente una de las razones por las que el gobierno podría querer llevar a cabo tal programa.

Los efectos de equilibrio ofrecen la única razón clara para favorecer estudios grandes sobre los pequeños. Esto no significa necesariamente regresiones de sección cruzada de países -que a menudo combinan demasiadas fuentes de variación diferentes como para ser útiles para hacer afirmaciones causales (Acemoglu y Johnson, 2007, sobre el impacto de la curación de enfermedades sobre el crecimiento económico de los países es una excelente excepción)- sino más bien microestudios que utilizan cambios de política a gran escala. Éstos no suelen ser aleatorios, pero a menudo siguen ofreciendo la oportunidad de ser cuidadosos con las cuestiones de causalidad y, al mismo tiempo, nos ayudan con respecto a los efectos de equilibrio porque muchos de los efectos de equilibrio se internalizan. Un buen ejemplo de este tipo de investigación es el trabajo de Hsieh y Urquiola (2006) que utilizan un diseño cuasi-experimental para argumentar que un programa de vales escolares chileno no condujo a una mejora general en la oferta de habilidades, aunque cambió los patrones de clasificación entre las escuelas. Otros estudios concebidos específicamente para evaluar los posibles efectos sobre el equilibrio del mercado de las políticas son los de Acemoglu y Angrist (2000) y Duflo (2004).

Es evidente que no siempre se dispone de la oportunidad de realizar estudios cuasi-experimentales de buena calidad y, en cualquier caso, es probable que valga la pena comprobar si los resultados son coherentes con las pruebas experimentales. Por ejemplo, en el caso de los vales, esperamos que los efectos de equilibrio atenúen la respuesta de la oferta y, por lo tanto, esperamos que los estudios cuasiexperimentales de mayor envergadura generen efectos menores que los experimentos. Si hallamos lo contrario, podríamos empezar a preocuparnos por si el estudio más grande es fiable o representativo. En este sentido, los estudios experimentales y no experimentales pueden ser complementos en lugar de sustitutos.

Otro enfoque es tratar de estimar directamente el tamaño del efecto de equilibrio utilizando el método experimental. En una investigación en curso, Kremer y Muralidharan estudian el efecto de un programa de vales utilizando una doble aleatorización: aleatorizan los pueblos en los que se entregan los vales así como quién recibe los vales dentro de un pueblo. Comparando las estimaciones que obtendrán de los dos tratamientos esperan inferir el tamaño del efecto de equilibrio. Por supuesto, esto sólo se refiere a un nivel de equilibrio: la gente puede trasladarse a la aldea desde fuera y salir de ella para encontrar trabajo; en este caso puede funcionar mejor estimar lo que ocurre con la oferta de educación que con el precio de las aptitudes, pero es claramente un comienzo importante.

Un enfoque relacionado es combinar los resultados de diferentes experimentos: Utilizar un experimento (o más plausiblemente, un cuasi experimento) para tratar de estimar la elasticidad de la demanda de aptitudes, otro para estimar la oferta de enseñanza de calidad y otro para estimar cuánto contribuyen los vales a la creación de aptitudes. Este es un estilo de trabajo que requiere adoptar un enfoque más estructural ya que necesitamos identificar cuáles son los parámetros relevantes. Como

discutiremos en la siguiente subsección, este tipo de trabajo está comenzando a realizarse, pero claramente hay un largo camino por recorrer.

2.5 Heterogeneidad de los Efectos de Tratamiento

La mayoría de las evaluaciones de los programas sociales se concentran exclusivamente en el impacto promedio. De hecho, una de las supuestas ventajas de los resultados experimentales es su simplicidad: son fáciles de interpretar, ya que lo único que hay que hacer es comparar promedios, lo que podría alentar a los encargados de la formulación de políticas a tomar más en serio los resultados (véase, por ejemplo, Duflo, 2004; Duflo y Kremer, 2004). Sin embargo, como señalan Heckman, Smith y Clements (1997), es posible que el efecto medio del tratamiento no sea lo que el formulador de políticas quiere saber: El enfoque exclusivo en la media sólo es válido bajo supuestos bastante específicos sobre la forma de la función de bienestar social. Además, desde el punto de vista del proyecto intelectual general, es evidente que no tiene sentido restringir el análisis a la comparación ingenua de promedios.

Lamentablemente, resulta que el efecto medio del tratamiento (o el efecto del tratamiento condicionado a las covariables) es también el único estadístico convencional de la distribución de los efectos del tratamiento fácil de estimar a partir de un experimento aleatorio sin hacer ningún otro supuesto adicional. Por supuesto, en principio, se podría comparar toda la distribución de resultados del tratamiento con la del control: existen pruebas de igualdad de distribuciones, así como de dominancia estocástica (véase Abadie, 2002). Por ejemplo, Banerjee, Cole, Duflo y Linden (2007) muestran que la distribución de resultados de las pruebas entre quienes estudian en escuelas que recibieron un Balsakhi el primer orden domina estocásticamente al del grupo de tratamiento, y la mayoría de las ganancias se observan en el extremo inferior. Esto es importante, ya que en las aulas del programa los niños de abajo fueron apartados y se les dio una enseñanza de recuperación, mientras que los de arriba permanecen en el aula. Por lo tanto, esperaríamos efectos muy diferentes en los dos grupos, y sería difícil justificar el programa si sólo ayuda a los de arriba. Duflo, Hanna y Ryan (2007) también analizan cómo el programa de incentivo para maestros basado en la cámara, del que se ha hablado anteriormente, afecta a toda la distribución de ausencias entre maestros, y descubren una dominancia estocástica de primer orden. Sin embargo, la comparación de estas distribuciones no nos informa sobre la distribución del efecto del tratamiento per se (ya que las diferencias en cuantiles de una distribución no es el cuantil de la diferencia).

En su excelente revisión de la reciente literatura econométrica sobre evaluación de programas (incluyendo detalles técnicos de gran parte del material aquí tratado), Imbens y Wooldridge (2008) argumentan que la distribución del resultado en tratamiento y en control (que siempre es conocido) es todo lo que podríamos querer saber sobre el programa, porque cualquier función de bienestar social debería definirse por la distribución de los resultados (o por la distribución de los resultados, condicionada a variables observables).

Sin embargo, no está claro que esto sea del todo correcto. Para ver el asunto en su forma más descarnada, consideremos el siguiente ejemplo. Hay una población de 3 personas, y sabemos sus posibles resultados si se trata y si no se trata. El resultado potencial del Sr. 1 si no se trata es 1, el del Sr. 2 es 2, y el del Sr. 3 es 3. El resultado potencial del Sr. 1 si se trata es 2, el del Sr. 2 es 3 y el del Sr. 3 es negativo 4. ¿Qué deberíamos pensar de este programa? Claramente, tanto en términos del efecto medio del tratamiento como en términos de la distribución global, el tratamiento fracasó: la distribución 1,2,3 del resultado potencial "no tratado" de primer orden domina la distribución -4, 2,3 del resultado potencial "tratado". ¿Debemos, por tanto, concluir que un responsable político debe favorecer siempre el control sobre el tratamiento? No necesariamente, ya que el tratamiento hace que la mayoría esté mejor y el legislador podría preocuparse por el "mayor bien del mayor número". Y aunque no estemos de acuerdo con las preferencias del responsable de las políticas en este caso, es difícil argumentar que el evaluador deba dictar la elección de la función objetivo.

Una vez que reconozcamos que podríamos preocuparnos por identificar el conjunto de personas (de un grupo indiferenciado ex ante) que subieron o bajaron debido al tratamiento, obviamente hay un problema. No hay manera de extraer esta información de la distribución de los resultados en el tratamiento y en el control, un hecho que está estrechamente relacionado con la observación de Heckman (1992) de que ni siquiera los experimentos producen efectos de tratamiento de cuantiles sin supuestos adicionales.

Se trata, por supuesto, de un problema lógico, y no de un problema de experimentos per se o de cualquier otra estrategia de estimación específica: simplemente no hay información relevante. En el marco de un experimento social aleatorio, Heckman, Smith y Clements (1997), muestran que al introducir supuestos de comportamiento adicionales (en efecto, modelando la decisión de participar en función de los resultados potenciales bajo tratamiento y sin tratamiento) se pueden estimar límites bastante precisos en las características de la distribución del efecto del tratamiento. Estas técnicas también se aplican en entornos no experimentales, pero los autores señalan que pueden ser especialmente útiles con los datos experimentales tanto porque "se puede hacer abstracción de los problemas de selección que afectan a los datos no experimentales", como porque el entorno experimental garantiza que haya un equilibrio en el apoyo de las variables observables, que es algo en lo que confían.

Nuestra opinión es que la investigación experimental ganaría algo si se comprometiera más con este grupo de investigaciones. Informar de algunos resultados más "dependientes de supuestos" junto con resultados más "libres de supuestos" que se suelen informar en los experimentos (y hacer la necesaria *caveat emptor*) sólo puede enriquecer el trabajo experimental. Sin embargo, los experimentos siguen teniendo ventaja sobre métodos que, con muy pocos supuestos, permiten conocer aspectos muy importantes de las repercusiones del tratamiento (como la media de cualquier subgrupo). El hecho de que podamos querer ir más allá de estas medidas,

y para ello podamos tener que invocar supuestos que puedan hacer que la asignación aleatoria sea menos importante, no puede contarse a favor de métodos no basados en una asignación aleatoria.

Además, gran parte de la heterogeneidad que caracteriza a las funciones objetivo de las personas (frente a la gran heterogeneidad que impulsa el resultado económico) no se refiere realmente a diferencias no observadas en características de las personas, sino a diferencias potencialmente observables. Por ejemplo, en el experimento del balsakhi (Banerjee, Cole, Duflo y Linden, 2007), no sólo observamos que la distribución de puntajes en el tratamiento de primer orden dominaba estocásticamente a la de control; también vimos que los que tenían puntajes bajos en la línea de base eran los que más ganaban. Desde el punto de vista de la organización ejecutora, Pratham, esto era lo que realmente importaba, pero sólo podíamos saberlo porque teníamos puntuaciones de prueba de referencia. En otras palabras, necesitamos comenzar el experimento con hipótesis claras sobre cómo varían los efectos del tratamiento en función de las covariables, y recopilar los datos de referencia pertinentes.

Afortunadamente, la reciente investigación econométrica puede ayudarnos mucho aquí. Crump y otros (2008), que ya hemos comentado anteriormente, desarrollan dos pruebas no paramétricas para determinar si existe heterogeneidad en los efectos de tratamiento: una para determinar si el efecto de tratamiento es nulo para cualquier subpoblación (definida por las covariables), y otra para determinar si el efecto de tratamiento es el mismo para todas las subpoblaciones (definidas por las covariables).

Además, se pueden estimar los efectos de tratamiento para diferentes subgrupos. Una dificultad en este caso es que si los subgrupos son determinados ex post, existe el peligro de "búsqueda de especificaciones", en la que los investigadores y los formuladores de políticas eligen ex post para enfatizar el impacto del programa en un subgrupo en particular. Aquí también, como en la aplicación de Heckman, Smith y Clements (1997), la teoría puede ayudar diciéndonos qué esperar. Especificar ex ante los resultados que se deben observar y lo que esperamos de ellos (como se alienta en la literatura médica) es otra posibilidad. Por supuesto que todavía podemos tratar de aprender de diferencias posiblemente interesantes (pero ex ante inesperadas) en el efecto del tratamiento. Este es otro lugar en el que la replicación puede ayudar: cuando se realiza un segundo experimento, éste puede configurarse explícitamente para probar esta hipótesis recién generada. Por ejemplo, Karlan y Zinman (2007) encuentran resultados muy diferentes para hombres y mujeres: los hombres están sometidos a un riesgo moral pero no a una selección muy adversa mientras que las mujeres lo están a la inversa. Estas diferencias no eran esperadas, y es difícil saber qué hacer con ellas. Pero una vez que el estudio se replique en otro lugar, éstas pueden constituir la base de un nuevo conjunto de hipótesis que se pondrá a prueba (véase Duflo, Kremer y Glennerster, 2008, para un análisis más detallado de estas y otras cuestiones de diseño).

Por último, una literatura reciente (Manski 2000, 2002, 2004, Dehejia, 2005, Hirano y Porter, 2005) trata de hacer todo esto menos ad hoc. Quieren integrar el proceso de evaluación y aprendizaje en un marco explícito de diseño de programas. Por lo tanto, tratan de ponerse explícitamente en la piel del formulador de políticas que trata de decidir si se aplica o no un programa, pero también cómo se aplica (¿debería ser obligatorio el programa? ¿Debería darse al administrador algún margen de maniobra sobre quién debe participar?). Permiten que el formulador de políticas se preocupe no necesariamente sólo por la ganancia de ingresos prevista, sino por la ganancia de utilidad prevista (teniendo en cuenta la aversión al riesgo) y, por lo tanto, por el posible aumento o disminución de la variabilidad del resultado con el estado del tratamiento. El formulador de políticas tiene acceso a covariables sobre los posibles beneficiarios, así como a resultados de experimentos aleatorios. En esta literatura se intenta desarrollar una teoría sobre cómo debe decidir el administrador, teniendo en cuenta tanto la heterogeneidad como la incertidumbre en los beneficios del programa condicionado a las covariables. Hasta donde sabemos, estas herramientas no han sido utilizadas en la investigación económica del desarrollo. Esta es una avenida fructífera para el trabajo futuro.

2.6 Relación con la Estimación Estructural

La mayor parte de la literatura experimental temprana se centraba en estimaciones de forma reducida del efecto del programa. Pero no hay razón para no utilizar también esos datos para extraer parámetros estructurales siempre que sea posible. Aunque esto requerirá que hagamos más supuestos, las estimaciones estructurales pueden utilizarse para comprobar los resultados de la forma reducida (¿son razonables los resultados si implican una elasticidad de la oferta de mano de obra de $\boldsymbol{x}$ o un rendimiento esperado de la escolarización de $\boldsymbol{y}$?) y más generalmente para reforzar su validez externa. Además, si nos sentimos cómodos con los supuestos que subyacen a las estimaciones, es posible derivar de ellas conclusiones de política que van mucho más allá de lo que se podría obtener de la forma reducida.

Entre los primeros ejemplos de este método figuran Attanasio, Meghir y Santiago (2002) y Todd y Wolpin (2006), que utilizan datos de PROGRESA. Attanasio, Meghir y Santiago están interesados en evaluar el impacto del programa, permitiendo, por ejemplo, efectos de anticipación en el control (lo que no puede hacerse sin formular algunos supuestos adicionales). No hallan ninguna evidencia de efectos de anticipación. Todd y Wolpin (2006) quieren utilizar el experimento como una forma de validar el modelo estructural: estiman un modelo estructural fuera de la muestra tratada y comprueban que el modelo predice correctamente el impacto del tratamiento. Otro ejemplo del potencial de casar los experimentos con la estimación estructural es el de Duflo, Hanna y Ryan (2007). Después de informar de los resultados de la forma reducida, el documento explota el aspecto no lineal de los planes de incentivos para maestros de Seva Mandir (los maestros recibían un salario mínimo de 10 dólares si estaban presentes menos de 10 días en el mes, y una bonificación de 1 dólar por cada día extra que superara esa cantidad) para estimar el valor

del maestro de no ir a la escuela y la elasticidad de su respuesta con respecto a la bonificación. El modelo es extremadamente simple (al venir a la escuela en los primeros días de los meses, el maestro está construyendo la opción de obtener un dólar extra al día al final, y renunciando a una opción externa estocástica de no ir ese día), pero da lugar a interesantes problemas de estimación, una vez que queremos introducir heterogeneidad y correlación serial en el shock recibido por el maestro en la opción externa de una manera realista. Al igual que Todd y Wolpin, este trabajo compara entonces las predicciones de varios modelos tanto con el control, como con un "experimento natural" en el que Seva Mandir cambió sus reglas de pago (después de que el período del experimento terminara). Este ejercicio demostró que es importante tener en cuenta tanto la heterogeneidad como la correlación serial, ya que sólo esas simulaciones se acercan a replicar los medios del grupo de control y la distribución de la ausencia según las nuevas reglas.

En principio, debería ser posible ir aún más lejos en la explotación de esa complementariedad entre estimación estructural y experimentos. Como ya se ha mencionado, una ventaja de los experimentos es la flexibilidad con respecto a la recopilación de datos y la elección de tratamientos: Debería ser posible diseñar el experimento para facilitar la estimación estructural asegurándose de que el experimento incluya fuentes de variación que ayuden a identificar los parámetros necesarios y a reunir el tipo de datos adecuado. También se podría estimar un modelo estructural a partir de datos de referencia antes de conocer los resultados del experimento, a fin de realizar una validación "a ciegas" de los modelos estructurales. Sin embargo, aún no se han visto ejemplos de este tipo de trabajo: los ejemplos que discutimos explotaron la variación ex post en la forma en que se implementó el programa, en lugar de introducirlo a propósito.

2.7 Relación con la teoría

Ya hemos argumentado que los experimentos pueden ser y han sido muy útiles para probar teorías (véase Banerjee (2005) y Duflo (2006) para un tratamiento más prolongado de estas cuestiones). El hecho de que los resultados experimentales básicos (por ejemplo, el efecto medio del tratamiento) no dependan de la teoría para su identificación, significa que puede ser posible una prueba "limpia" de la teoría (es decir, una prueba que no dependa también de otras teorías).

Un punto en el que esto ha sido claramente muy útil es en hacernos repensar algunos elementos básicos de teoría de la demanda. Un hallazgo consistente de una serie de estudios aleatorios independientes de demanda de lo que podría llamarse productos de protección de la salud, es que la elasticidad precio de la demanda en torno a cero es enorme. Kremer y Miguel (2007) comprobaron que al aumentar el precio de los medicamentos antiparasitarios de 0 a 30 centavos por niño en Kenya se redujo la fracción de niños que tomaban el medicamento del 75% al 19%. También en Kenya, Cohen y Dupas (2007) observaron que al aumentar el precio de los mosquiteros tratados con insecticidas de 0 a 60 centavos se reduce la fracción de los que compran

esos mosquiteros en 60 puntos porcentuales. El aumento del precio del desinfectante de agua de 9 centavos a 24 centavos reduce en 30 puntos porcentuales la fracción que acepta la oferta en Zambia (Ashraf, Berry y Shapiro, 2007). También se encuentran respuestas similares de gran envergadura con pequeños subsidios: En la India, Banerjee, Duflo, Glennerster y Kothari (2008) observaron que ofrecer a las madres un kilo de frijoles secos (por un valor de unos 60 centavos) por cada visita de vacunación (más un juego de tazones para completar la vacunación) aumenta en 20 puntos porcentuales la probabilidad de que un niño esté totalmente inmunizado. Y lo más notable es que un premio de 10 centavos hizo que un 20% más de personas en Malawi recojan los resultados de su prueba de VIH (Thornton, 2008).

Kremer y Holla (2008), al examinar esta evidencia (y varios documentos sobre educación con conclusiones similares), concluyen que estas elasticidades de demanda no pueden provenir del modelo estándar de capital humano de demanda por mejor salud, dada la importancia del tema en cuestión. Por ejemplo, se puede imaginar que un agente económico convencionalmente racional podría decidir someterse a la prueba de VIH (conocer su situación podría prolongar su vida y la de los demás) o podría decidir no someterse a ella (la prueba puede ser extremadamente estresante y vergonzosa). Lo que es más difícil de entender es que muchos de ellos cambien de opinión por sólo 10 centavos, sobre algo que tiene muchas posibilidades de transformar completamente sus vidas, de una manera u otra.

Kremer y Holla (2008) sugieren que esta pauta de demanda es más coherente con un modelo en el que la gente quiere realmente el producto pero lo aplaza; es tentador retrasar el pago del costo, dado que los beneficios están en el futuro. Por otra parte, si es cierto que la gente realmente quiere comprar mosquiteros o conocer el resultado de sus pruebas pero es permanentemente incapaz de hacerlo, entonces, dados los posibles beneficios que ofrecen para salvar vidas, tienen que ser extraordinariamente ingenuos. Sin embargo, cuando se trata de productos financieros, las pruebas (experimentales) se oponen a que sean tan ingenuos. Ashraf, Karlan y Yin (2007) observan que quienes muestran preferencias particularmente hiperbólicas son los que están más dispuestos a adquirir dispositivos de compromiso para fijar sus ahorros, lo que indica un alto grado de autoconciencia. Duflo, Kremer y Robinson (2008) observan que los agricultores de Kenya que se quejan de no tener suficiente dinero para comprar fertilizantes en el momento de la siembra están dispuestos a comprometer dinero en el momento de la cosecha para que el fertilizante se utilice en la siembra varios meses después. Además, cuando se les da ex ante (antes de la cosecha) la opción de decidir cuándo deben venir a vender fertilizantes, casi la mitad de los agricultores les piden que vengan justo después de la cosecha, en lugar de más tarde, cuando necesitarán fertilizante, porque saben que tendrán dinero después de la cosecha. Sin embargo, piden que se les entregue el fertilizante de inmediato, lo que sugiere que tienen al menos suficiente autocontrol para guardar el fertilizante en casa y no revenderlo. Esto sugiere que la teoría podría ir más allá de la invocación, ahora estándar, de problemas de autocontrol como una forma de tratar todas las anomalías.

A veces los experimentos arrojan resultados que son aún más preocupantes para el conjunto de la teoría existente (véase Duflo (2004) para una discusión más larga). Un ejemplo sorprendente que no encaja con ninguna teoría económica existente es el de Bertrand, Karlan, Mullainathan, Shafir y Zinman (2008): Observan que manipulaciones aparentemente menores (como la fotografía en un gestor de correo) tienen efectos sobre la aceptación de préstamos tan grandes como cambios significativos en los tipos de interés.

En todo esto, los experimentos juegan el papel que tradicionalmente desempeñan los experimentos de laboratorio, quizás con mayor credibilidad. El objetivo es una mejor teoría. ¿Pero puede la teoría ayudarnos a diseñar mejores experimentos e interpretar mejor los resultados experimentales para un mejor diseño de políticas? Una posible dirección, discutida anteriormente, es usar resultados experimentales para estimar modelos estructurales. Sin embargo, también queremos que la teoría juegue un papel más mundano pero igualmente importante: Necesitamos un marco para interpretar lo que encontramos. Por ejemplo, ¿podemos ir más allá de la observación de que los diferentes insumos de la función de producción educativa tienen diferentes productividades? ¿Hay alguna manera de agrupar los diferentes insumos en categorías de insumos más amplias a priori, con la presunción de que debería haber menos variación dentro de la categoría? O por el lado de los resultados, ¿podemos predecir qué resultados del sistema educativo deberían co-moverse mucho más estrechamente que el resto? ¿O cada resultado experimental es sui generis?

Es poco probable que la teoría que sería útil para este propósito sea particularmente sofisticada. Más bien, al igual que el famoso modelo de Mincer, sería una forma conveniente de reducir la dimensionalidad, basada en un conjunto de algunas premisas razonables. Banerjee y otros (2008) intentan hacer algo así para el caso de la acción pública local, pero su esfuerzo es, en el mejor de los casos, parcialmente exitoso. Será vital trabajar más en esta línea.

3. Conclusión

Por lo tanto, estamos totalmente de acuerdo con el punto principal de Heckman (1992): para ser interesantes, los experimentos deben ser ambiciosos, y deben estar informados por la teoría. Es también, convenientemente, donde es probable que sean más útiles para los encargados de la formulación de políticas. Nuestra opinión es que las ideas de los economistas pueden y deben guiar la elaboración de políticas (véase también Banerjee, 2002). A veces están bien situados para proponer o identificar programas que probablemente marquen grandes diferencias. Y lo que es aún más importante, a menudo están en condiciones de poner en marcha el proceso de descubrimiento de políticas, basado en la interacción de teoría e investigación experimental. Este proceso de "experimentación creativa", en el que responsables políticos e investigadores trabajan juntos para pensar de forma innovadora y aprender de los éxitos y los fracasos, es la contribución más valiosa del reciente aumento de la labor experimental en la economía.

Bibliografía

Abadie, Alberto (2002). "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *Journal of the American Statistical Association*, 97(457) 284-292.

Abdul Latif Jameel Poverty Action Lab, "Fighting Poverty: What Works?" Issue One, Fall 2005.

Acemoglu, Daron, and Joshua Angrist (2000). "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," NBER Macroeconomics Annual, Volume 15, pp. 9-59.

Acemoglu, Daron, and Simon Johnson (2007). "Disease and Development: the Effect of Life Expectancy on Economic Growth," *Journal of Political Economy*, December, Volume 115, pp. 925-985.

Angrist, Joshua, Eric Bettinger, and Michael Kremer, (2006). "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*. Volume 96(3), pp. 847-862.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Michael Kremer and Elizabeth King (2002). "Vouchers for Private Schooling in Colombia: Evidence from Randomized Natural Experiments," *The American Economic Review*, December, Volume 92(5), pp.1535-1558.

Angrist, Joshua, D. Lang, and Philip Oreopoulos (2009). "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*.

Angrist, Joshua and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, Volume 114(2) pp. 533-575.

Angrist, Joshua. and Victor Lavy (2002). "The Effect of High School Matriculation Awards: Evidence from Group-Level Randomized Trials," NBER Working Paper No. 9389.

Ashraf, Nava, Dean Karlan, and Wesley Yin (2006). "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics* 121(2), pp. 635-672.

Ashraf, Nava, James Berry and Jesse M. Shapiro (2007). "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," NBER Working Paper No. 13247.

Attanasio, Orazio, Costas Meghir and Ana Santiago (2001), "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa," UCL Mimeo.

Banaji, Mahzarin. (2001). 'Implicit attitudes can be measured," In *The nature of remembering: Essays in honor of Robert G. Crowder*, H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant editors, Washington, DC: American Psychological Association.

Banerjee, Abhijit (2002). "The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results," BREAD Working Paper No. 007.

Banerjee, Abhijit (2005) "'New Development Economics' and the Challenge to Theory," *Economic and Political Weekly*, Vol. 40(40), October 1-7, pp. 4340-4344.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, Stuti Khemani (2008) "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," MIMEO, MIT, 2008.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden (2007). "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, Volume 122(3), pp.1235-1264.

Banerjee, Abhijit., Esther Duflo, Rachel Glennerster and Dhruva Kothari (2008). "Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives," MIMEO MIT.

Banerjee, Abhijit, Suraj Jacob and Michael Kremer, with Jenny Lanjouw and Peter Lanjouw (2005). "Moving to Universal Education! Costs and Trade offs," MIMEO, MIT.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande and Petia Topalova (2008). "Powerful Women: Does Exposure Reduce Bias?" BREAD Working Paper No. 181, NBER working paper number 14198.

Berry, James (2008). "Rotten Kids or Rotten Parents: Child Motivation and Education Decision in India," MIMEO, MIT.

Bertrand, Marianne, Dolly Chugh and Sendhil Mullainathan (2005). "Implicit Discrimination," *American Economic Review*, Volume 95(2), pp. 94-98.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan (2009). "Corruption in Driving Licensing Process in Delhi," *Quarterly Journal of Economics*.

Bjorkman, Martina and Jakob Svensson (2007). "Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda," Community-Based Monitoring of Primary Health Care PCEPR Working Paper No. 6344.

Bleakley, Hoyt (2007). "Disease and Development: Evidence from Hookworm Eradication in the American South," *Quarterly Journal of Economics*, Volume 122(1), pp. 73-117. Blundell, Whitney Newey, Torsten. Persson, editors, Cambridge University Press, Vol. 2(42) (see also BREAD Policy Paper No. 002, 2005).

Bobonis, Gustavo, Edward Miguel, Charu Puri Sharma (2006). "Anemia and School Participation," *Journal of Human Resources*, Volume 41 (4), pp.692–721.

Chin, Aimee (2005). "Can Redistributing Teachers Across Schools Raise Educational Attainment? Evidence from Operation Blackboard in India," *Journal of Development Economics* 78, pp. 384-405.

Cull, Robert, David McKenzie and Christopher Woodruff (2007). "Experimental Evidence on Returns to Capital and Access to Finance in Mexico," *World Bank Economic Review*.

Cohen, Jessica and Pascaline Dupas (2007). "Free Distribution or Cost-Sharing? Evidence from a randomized malaria prevention experiment," *Brookings Institution Global Working Paper* No.14.

Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik, (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*.

Dehejia, Rajeev (2005). "Program Evaluation as a Decision Problem," *Journal of Econometrics*, Volume 125, pp. 141-173.

Duflo, Esther (2004a). "The Medium Run Consequences of Educational Expansion: Evidence from a Large School Construction Program in Indonesia," *Journal of Development Economics* Volume 74(1) pp. 163-197.

Duflo, Esther (2004b). "Scaling Up and Evaluation," in *Accelerating Development*, Francois Bourguignon and Boris Pleskovic, editors, World Bank and Oxford University Press: Washington, DC and Oxford, 2004, pp. 342-367.

Duflo, Esther (2006). "Field Experiments in Development Economics," in *Advances in Economic Theory and Econometrics,* Richard Blundell, Whitney Newey, Torsten Persson, editors, Cambridge University Press, Volume 2(42), see also BREAD Policy Paper No. 002, 2005.

Duflo, Esther and Raghabendra Chattopadhyay (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica*, Volume 72(5), pp.1409-1443.

Duflo, Esther, Pascaline Dupas and Michael Kremer (2008). "Peer Effects, Pupil Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya," MIMEO, MIT.

Duflo, Esther, Pascaline Dupas, Michael Kremer, and Samuel Sinei (2006). "Education and HIV/AIDS Prevention: Evidence from a randomized evaluation in Western Kenya," World Bank Policy Research Working Paper No.402.

Duflo, Esther, Rema Hanna, and Stephen Ryan (2007) "Monitoring Works: Getting Teachers to Come to School," NBER Working Paper No. 11880, 2005; BREAD Working Paper No. 103.

Duflo, Esther and Michael Kremer (2004). "Use of Randomization in the Evaluation of Development Effectiveness," in *Evaluating Development Effectiveness* (World Bank Series on Evaluation and Development, Volume 7, Osvaldo Feinstein, Gregory K. Ingram and George K. Pitman, editors, Transaction Publishers: New Brunswick, NJ, 2004, pp. 205-232.

Duflo, Esther, Michael Kremer, and Rachel Glennerster "Using Randomization in Development Economics Research: A Toolkit," in Handbook of Development Economics. Elsevier-North Holland John Strauss and Paul Schultz, editors, Volume 4.

Duflo, Esther, Michael Kremer and Jonathan Robinson (2008a). "How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya," *American Economic Review Papers and Proceedings*, Volume 98(2), pp. 482-488.

Duflo, Esther, Michael Kremer and Jonathan Robinson (2008b). "Why are Farmers not using Fertilizer? Procrastination and Learning in Technology adoption," MIMEO, MIT.

Dupas, Pascaline (2007). "Relative Risks and the Market for Sex: Teenage Pregnancy, HIV, and Partner Selection in Kenya," MIMEO, UCLA.

Gine, Xavier, Dean Karlan and Jonathan Zinman (2008). "Put Your Money Where Your Butt Is: A Commitment Savings Account for Smoking Cessation," MIMEO, Yale University.

Glewwe, Paul, Nauman Ilias, and Michael Kremer (2003). "Teacher Incentives". MIMEO, Harvard.

Glewwe, Paul, Michael Kremer and Sylvie (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal, Applied Economics*.

Glewwe, Paul, Michael Kremer, Sylvie, and E. Zitzewitz (2004). "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economic*. Volume 74(1), pp. 251-268.

Heckman, James J. (1992). "Randomization and social policy evaluation," in *Evaluating Welfare and Training Programs*, editors Charles Manski and I. Garfinkel. Cambridge, MA: Harvard University Press. (also available as NBER Technical Working Paper No.107, 1991).

Heckman, James J., Hidehiko Ichimura, J. Smith, and Petra Todd, (1998). "Characterizing Selection Bias Using Experimental Data," *Econometrica* Volume 66, pp. 1017-1098.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, Volume 64, pp.605-654.

Heckman, James J., Lance Lochner, and Christopher Taber, (1999). "Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy," *Fiscal Studies* Volume 20(1), pp. 25-40.

Heckman, James J., Jeffrey Smith, and Nancy Clements, (1997). "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies*, Volume 64, pp.487-535.

Hirano, Keisuke, and Jack Porter (2005). "Asymptotics for Statistical Decision Rules" *Econometrica*. Volume 71(5), pp. 1307-1338.

Holla, Alaka and Kremer, Michael (2008). "Pricing and Access: Lessons from Randomized Evaluation in Education and Health," MIMEO, Harvard University.

Hsieh, Chang-Tai and Miguel Urquiola (2006). "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program," *Journal of Public Economics*. Volume 90, pp.1477–1503.

Imbens, Guido, and Joshua Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Volume 61(2), pp. 467-476.

Imbens, Guido and Jeffrey M. Wooldridge (2008). "Recent Developments in the Econometrics of Program Evaluation," MIMEO, Harvard University (*Journal of Economic Literature*, 2009).

Karlan, Dean and Jonathan Zinman (2005). "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," BREAD Working Paper No. 94.

Karlan, Dean (2005a). "Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions," *American Economic Review*, Volume 95(5), pp. 1688-1699.

Karlan, Dean, and Jonathan Zinman (2007). "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts," MIMEO, Yale University.

Karlan, Dean and Jonathan Zinman (2008). "Credit Elasticities in Less Developed Countries: Implications for Microfinance," *American Economic Review*, Volume 98(3), pp.1040-1068.

Kremer, Michael, and Edward Miguel (2007). "The Illusion of Sustainability," *Quarterly Journal of Economics*, Volume 122(3), pp. 1007-1065.

Kremer, Michael, Edward Miguel, and Rebecca Thornton (2009). "Incentives to Learn," *Review of Economics and Statistics*, see also NBER Working Paper No. 10971 (2007).

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Zwane. "Spring Cleaning: Rural Water Impacts, Valuation, and Institutions," MIMEO, Berkeley.

Manski, Charles, (2000). "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice". *Journal of Econometrics*, Volume 95, pp. 415-442.

Manski, Charles, (2002). "Treatment Choice Under Ambiguity Induced by Inferential Problems," *Journal of Statistical Planning and Inference*, Volume 105, pp. 67-82.

Manski, Charles, (2004). "Measuring Expectations," *Econometrica*, Volume 72(4), pp. 1329-1376.

Manski, Charles, (2004). "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, Volume 2(4), pp. 1221-1246.

McKenzie, David, Suresh de Mel, and Christopher Woodruff (2008). "Returns to Capital: Results from a Randomized Experiment," *Quarterly Journal of Economics*.

Miguel, Edward and Michael Kremer (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, Volume 72 (1), pp. 159-217.

Olken, Benjamin (2007). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, Volume 115 (2), pp. 200-249.

Olken, Benjamin, and Patrick Barron (2007b). "The Simple Economics of Extortion: Evidence from Trucking in Aceh," NBER Working Paper No. 13145, BREAD Working Paper No. 151, *CEPR Discussion Paper* No. 6332.

Rodrik, Dani (2008). "The New Development Economics: We Shall Experiment, But How Shall We Learn?" MIMEO, Harvard University.

Rubin, Donald, (2006). "*Matched Sampling for Causal Effects*," Cambridge University Press, Cambridge, UK.

Thornton, Rebecca, (2008). "The Demand for and Impact of HIV Testing: Evidence from a Field Experiment," *American Economic Review*.

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