J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

INTRODUCCION A LAS EVALUACIONES

Una Evaluación Aleatoria es un tipo de Evaluación de Impacto que usa un proceso aleatorio para asignar recursos, ejecutar programas, o para aplicar políticas como parte del diseño del estudio. Como todas las evaluaciones de impacto, el propósito principal de las evaluaciones aleatorias es el de determinar si un programa tiene impacto, y más específicamente, cuantificar la magnitud del impacto. Las evaluaciones de impacto típicamente miden la efectividad de un programa al comparar los resultados de aquellos (individuos, comunidades, escuelas, etc.) que recibieron el programa, frente a aquellos que no. Hay varios métodos para hacer esto, pero las evaluaciones aleatorias son generalmente consideradas las más rigurosas y, con todo lo demás constante, producen los resultados más precisos (es decir, sin sesgo).

La sección de Metodología cubre el **qué, por qué, quién, cuándo y cómo** de las evaluaciones aleatorias.

Para más información acerca de evaluaciones aleatorias, visite:

- **Running Randomized Evaluations**
  R. Glennerster and K. Takavarasha, November 2013
- **Field Experiments: Design, Analysis and Interpretation**
  A. Gerber and D. Green, May 2012
- **Evaluando Programas Sociales: Curso** Ejecutivo del Poverty Action Lab
- Una **versión en línea gratuita del curso**
- **Using Randomization in Development Economics Research: A Toolkit**. E. Duflo, M. Kremer y R. Glennerster
- **Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons.** M. Kremer
- **Field Experiments in Development Economics.** E. Duflo, Enero de 2006
- **Use of Randomization in the Evaluation of Development Effectiveness.** E. Duflo y M. Kremer, Julio de 2003
- **Scaling Up and Evaluation.** E. Duflo, Mayo de 2003
- **Nonexperimental Versus Experimental Estimates of Earnings Impacts** S. Glazerman, D. Levy y D. Myers, Mayo de 2003.
- **Impact Evaluation in Practice**
  P. Gertler, S. Martinez, P. Premand, L. Rawlings and C. Vermeersh

TABLE OF CONTENTS

¿QUÉ ES UNA EVALUACIÓN?

La palabra "evaluación" puede ser interpretada de manera bastante amplia. Significa cosas distintas para distintas personas y organizaciones. Los ingenieros, por ejemplo, pueden evaluar o probar la calidad del diseño de un producto, la durabilidad del material, la eficiencia de un proceso productivo o la seguridad de un puente. Los críticos evalúan o reseñan la calidad de un restaurant, película o libro. Un psicólogo de niños puede evaluar o valorar el proceso de decisión de los niños.

Los investigadores en J-PAL evalúan programas sociales y políticas públicas diseñadas para mejorar el bienestar de las personas pobres del mundo. Esto se conoce como evaluación de programas.

En pocas palabras, la evaluación de un programa esta destinada a responder la pregunta: "¿Cómo está funcionando nuestro programa o política?". Esto puede tener distintas respuestas dependiendo de quién esté preguntando, y a quién le están hablando. Por ejemplo, si un donante pregunta al Director de la ONG "¿Cómo está funcionando nuestro programa?" esto puede implicar: "¿Has estado malgastando nuestro dinero?" Eso puede sentirse como un interrogatorio. Alternativamente, si un político pregunta a su electorado, "¿Cómo está funcionando nuestro programa?", podría estar simplemente preguntando: "¿Está nuestro programa alcanzando sus metas? ¿Cómo podemos mejorarlo para usted?". Por ende, la evaluación de programas puede ser asociada con sentimientos positivos o negativos, dependiendo si su objetivo es el de exigir una rendición de cuentas o si se trata de un deseo de aprender.

J-PAL trabaja con gobiernos, ONGs, donantes, y otros socios que están más interesados en aprender las respuestas a preguntas como: ¿Cuán efectivo es nuestro programa? Esta respuesta puede ser dada a través de una evaluación de impacto. Hay varios métodos para realizar **evaluaciones de impacto**, pero la que usa J-PAL es la **evaluación aleatoria**.

A un nivel muy básico, las evaluaciones aleatorias pueden responder la pregunta: ¿Fue efectivo el programa? Pero si está bien pensado su diseño e implementación, también puede responder a las preguntas: ¿Cuán efectivo fue? ¿Hubo efectos involuntarios? ¿Quién se beneficio más? ¿Quién salió perjudicado? ¿Por qué funcionó o por qué no? ¿Qué aprendizajes pueden ser aplicados en otros contextos, o si el programa se lleva a mayor escala? ¿Cuán costo-efectivo resultó el programa? ¿Cómo se compara con otros programas diseñados para cumplir los mismos objetivos? Para responder estas (tan interesantes, si no es que más interesantes) preguntas, el programa de evaluación debería ser parte de un paquete más grande de evaluaciones y ejercicios. Siguiendo el marco de "Comprehensive evaluations" de Rossi, Freeman y Lipsy, este paquete será cubierto en las siguientes secciones:

1. **Evaluación de Necesidades**
2. **Evaluación Teórica del Programa**
3. **Evaluación de Procesos**
4. **Evaluación de Impacto**
5. **Análisis de Costo-beneficio, Costo-efectividad, y Costo-comparación**
6. **Objetivos, Resultados y Mediciones**

Las primeras dos evaluaciones (**Necesidades** y **Teoría del Programa**) se refieren a las necesidades que busca cubrir este programa y cuáles son los pasos mediante los cuales logrará estos objetivos. Idealmente estos pasos deberían ser fijados por las personas que llevarán a cabo la implementación, antes de que se establezca la evaluación de impacto.

Las **evaluaciones de procesos** son útiles para los administradores del programa y para medir si los hitos y resultados se están logrando a tiempo. Muchas organizaciones han establecido sistemas de seguimiento del proceso – a menudo clasificados como Evaluación y Monitoreo (E&M).

Las **evaluaciones de impacto** están diseñadas para medir si el programa o la política están teniendo éxito en el logro de sus objetivos.

Finalmente, los **análisis costo-beneficio** y **costo-efectividad** son útiles para las implicancias políticas de un programa. El primero observa si los beneficios alcanzados por el programa justifican su costo. El segundo compara los beneficios de este programa frente a otros programas diseñados para lograr objetivos similares.

En la realización de cualquier análisis o evaluación es imperativo pensar acerca de cómo se puede medir el progreso. Los indicadores de progreso – manteniendo las metas de los programas y los resultados esperados en mente – requieren una reflexión importante así como también un sistema de recolección de datos. Esto se cubre en **Objetivos, resultados y mediciones.**

EVALUACION DE NECESIDADES

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

Por ejemplo, el problema podría deberse al escurrimiento de fertilizantes orgánicos que están contaminando el agua que beben ciertas comunidades.

La evaluación de necesidades es un enfoque sistemático para identificar la naturaleza y el alcance de un problema social, definir la población objetivo a ser atendida, y determinar la atención que necesitan para hacer frente al problema.

Una evaluación de necesidades es esencial, porque los programas serán inefectivos si el servicio no se diseña adecuadamente para atender las necesidades o si las necesidades realmente no existen. Por ejemplo, si las fuentes que contaminan el agua potable están relacionadas con la agricultura, las inversiones en infraestructura de saneamiento, tales como baños y sistemas de alcantarillado, podrían no resolver el problema. La evaluación de necesidades puede ser conducida utilizando indicadores sociales, encuestas y censos, entrevistas, etc.

EVALUACIÓN TEÓRICA DEL PROGRAMA

Los programas y políticas se realizan para enfrentar necesidades específicas. Encontrar esa necesidad, usualmente, requiere más reflexión que el encontrar y presionar un botón o tomar una píldora. Para los responsables de hacer políticas públicas, requiere la identificación de las razones que causan esos resultados indeseables (ver **evaluación de necesidades**), y elegir estrategias de una larga lista de opciones para lograr tratar de tener distintos resultados.

Por ejemplo, si las personas están tomando agua contaminada, un programa podría ser diseñado para prevenir que el agua sea contaminada – mejorando la infraestructura de saneamiento – mientras que otra podría ser diseñada para tratar el agua contaminada utilizando cloro. Una propuesta de intervención podría tener como objetivo a aquellos responsables de la contaminación, otra podría apuntar a los que toman el agua. Una estrategia podría descansar en el supuesto de que las personas no saben que el agua está sucia, otra, que ellos saben pero no tienen acceso a cloro, e incluso otra, sería que aún sabiendo y teniendo el acceso a cloro, no lo hacen porque tienen otras razones (por ejemplo, falta de información, sabor, costo, etc.). Estos programas deben analizar simultáneamente las restricciones de capacidades (financieras, humanas e institucionales) y las realidades políticas de sus contextos. Al concebir una respuesta apropiada, los actores de políticas públicas, implícitamente, toman decisiones acerca de cuál es la mejor intervención y por qué. Cuando este ejercicio mental es documentado explícitamente de forma estructurada, los responsables de hacer política pública están conduciendo lo que se conoce como *evaluación teórica del programa*, o *evaluación de diseño*.

Una Evaluación Teórica del Programa modela la teoría que está detrás del programa, presentando un plan viable y factible para mejorar la condición social del objetivo. Si las metas y supuestos son irracionales, entonces hay pocas posibilidades de que el

programa sea efectivo. La evaluación teórica del programa incluye primero, articular el programa teórico y después evaluar cuán bien la teoría responde a las necesidades de la población objetivo. Las metodologías usadas en la evaluación teórica de programas incluyen el *Enfoque del Marco Lógico* o **Teoría del Cambio**.

La siguiente tabla es un ejemplo simple de un marco lógico:

Needs	Input	Output	Outcome	Impact	Long-Term Goal
People are frequently sick from drinking contaminated water and do not currently use methods to treat their water	NGO purchases chlorine tablets and develops infrastructure for distribution to households	Households receive chlorine tablets	Individuals stop drinking contaminated water and start drinking treated water	Incidence of diarrhea decreases	Decrease in mortality, particularly child mortality. Improved physical and cognitive development

EVALUACIÓN DE PROCESOS

Antes de ser lanzado, cualquier programa existe a nivel conceptual – como un diseño, descripción o plan (vea **Evaluación Teórica del Programa**). Pero una vez lanzando, el programa enfrenta realidades de terreno: ¿La organización cuenta con un buen y entrenado equipo de trabajo? ¿Están las responsabilidades bien asignadas? ¿Están siendo completadas las tareas de los intermediarios a tiempo? Si el programa fue diseñado para proveer tabletas de cloro a los hogares para tratar el agua contaminada, por ejemplo, ¿Están alcanzando a entregar la cantidad apropiada de tabletas de cloro en los centros de distribución a tiempo?

La Evaluación de procesos, también conocida como *evaluación de la implementación* o *evaluación del proceso del programa*, analiza la efectividad de las operaciones del programa, la implementación y la entrega de servicios. Cuando la evaluación de procesos está en curso se llama *monitoreo del programa* (como en Evaluación y Monitoreo: E&M). La evaluación de procesos nos ayuda a determinar, por ejemplo:

- Si los servicios y metas están alineados apropiadamente.
- Si los servicios están siendo entregados a los destinatarios, como se pretendía.
- Cuán bien está organizado el servicio de entrega.
- La efectividad de la gestión del programa.
- Cuán efectivamente se están usando los recursos del programa.[1]

Las evaluaciones de procesos son usadas a menudo por los administradores como puntos de referencia para medir el éxito, por ejemplo: la distribución de tabletas de cloro está alcanzando el 80% de los beneficiarios que se pretendían por semana. Estos puntos de referencia pueden ser fijados por administradores del programa, y a veces por donantes. En muchas organizaciones grandes, la supervisión del progreso es la responsabilidad de un departamento de Evaluación y Monitoreo (E&M). Con el fin de determinar si se están alcanzando los puntos de referencia, mecanismos de **recolección de datos** deben existir.

[1] Rossi, Peter, et al. Evaluation. A Systematic Approach. Thousand Oaks: Sage Publications, 1999.

EVALUACIÓN DE IMPACTO

Los programas y las políticas están diseñados para alcanzar una meta (o una serie de metas). Por ejemplo, un programa para la distribución de cloro puede ser implementado específicamente para combatir la alta incidencia de enfermedades transmitidas por el agua en una región. Podríamos preguntarnos si el programa está resultando exitoso en lograr esta meta. Esto no es lo mismo que preguntar "¿El cloro mata la bacteria?" o "¿El consumo de cloro es perjudicial?". Estas preguntas pueden responderse en un laboratorio real. Para que nuestro programa alcance su meta de detener las enfermedades, se debe asignar el dinero, se deben comprar las tabletas de cloro, se deben acomodar los mecanismos de distribución, los hogares deben recibir las tabletas, deben usarlas, y no deben consumir agua no tratada. Una evaluación de programa nos ayuda a determinar si todos estos requisitos se están cumpliendo, y si nuestro objetivo se está logrando según lo previsto.

Como parte normal de la operación, ej. contabilidad básica, cierta información es producida, como cuantas cajas de tabletas de cloro han sido enviadas. Esto puede ser usado para la **evaluación de procesos**. Pero no nos puede decir si hemos reducido exitosamente la incidencia de diarrea. Para medir el impacto, debemos utilizar indicadores más directos, tales como el número de personas que declaró sufrir de diarrea en los últimos dos meses.

Las evaluaciones de impacto miden el éxito de un programa – donde el éxito puede ser una definición amplia o estrecha. Nos ayuda a eliminar las intervenciones menos eficaces de todas las intervenciones exitosas y mejorar los programas existentes.

Evaluación de impacto

El principal propósito de una evaluación de impacto es la determinar si un programa tiene impacto (en unos cuantos resultados clave), y más específicamente, cuantificar cuán grande es el impacto. ¿Qué es impacto? En nuestro ejemplo del cloro, impacto es cuánto más saludable están las personas gracias al programa de lo que podrían haber estado sin el mismo. O más específicamente, cuanto más disminuyo la incidencia de diarrea con el programa que sin éste.

Conseguir esta cifra correcta es más difícil de lo que parece. Es posible medir la incidencia de la diarrea en una población que recibe el programa, pero es imposible medir directamente "¿Cómo estarían si no hubiesen recibido el programa?" – así como es imposible medir cómo estaría la economía Estadounidense hoy si los Nazis hubiesen ganado la Segunda Guerra Mundial, o cual sería la enfermedad más mortal hoy en día si no se hubiese descubierto la penicilina en el sucio laboratorio de Alexander Fleming en 1928 en Londres. Es posible que Alemania se hubiese convertido en la economía dominante del mundo, o alternativamente, que los Nazis hubiesen caído unos años después. Es posible que pequeñas heridas siguieran siendo causantes de muchas muertes, o alternativamente, algo parecido a la penicilina hubiese sido descubierto en un laboratorio diferente en otra parte del mundo. En nuestro ejemplo de las tabletas de cloro, es posible que sin el cloro, las personas se hubiesen mantenido enfermas como lo estaban antes, o es posible que hubiesen empezado a hervir el agua – y que las tabletas de cloro sólo iban a servir como sustituto de una tecnología por otra – sugiriendo que las personas no están más saludables gracias a las tabletas de cloro.

Las evaluaciones de impacto, usualmente, estiman la efectividad de un programa al comparar los resultados de aquellos (individuos, comunidades, escuelas, etc.) que participaron en el programa frente a los que no lo hicieron. El desafío clave en una evaluación de impacto es *el encontrar un grupo de personas que no participaron*, pero que son lo suficientemente parecidas como para medir "*cómo estarían los participantes si no hubiesen recibido el programa*". Hay varios métodos para hacer esto y cada método viene acompañado de sus propios supuestos.

Una tabla comparando las diferentes metodologías se puede encontrar en la sección¿ **Por qué aleatorizar**?

ANÁLISIS DE COSTO-BENEFICIO/EFECTIVIDAD/COMPARACIÓN

Dos organizaciones pueden tener estrategias muy distintas para enfrentar el mismo problema. Si el suministro de agua de una comunidad, por ejemplo, fuera contaminado llevando a una gran epidemia de diarrea, una ONG puede abogar por realizar inversiones en infraestructura moderna para sanear el agua, incluyendo un sistema de alcantarillado, tuberías de agua, etc. Otra ONG podría proponer un sistema de distribución donde los hogares reciban, gratuitamente, tabletas de cloro para tratar el agua en su propia casa. Si estos dos métodos fuesen igualmente efectivos – cada uno reduciendo la diarrea en 80 porciento, ¿Estarían los responsables de hacer políticas públicas igual de contentos implementando una u otra? Probablemente no; ya que necesitarían considerar los costos de cada estrategia.

Es muy probable que la inversión en infraestructura moderna en un pueblo lejano sea prohibitivamente cara. En este caso, la opción sería clara. No obstante, las opciones no son siempre tan blancas o negras. Una opción más realista (pero aún hipotética) sería entre una inversión en infraestructura que reduce la diarrea en un 80 por ciento, frente a un programa de distribución de tabletas de cloro que cuesta 1/100ª parte del precio, y reduce la diarrea en un 50 porciento.

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

Un análisis de costo-efectividad toma el impacto de un programa (por ejemplo, porcentaje de reducción en la incidencia de la diarrea), y lo divide por el costo del programa, generado estadísticas tales como: el número de casos de diarrea prevenidos por dólar invertido. Esto no crea ningún juicio respecto del valor de la reducción de la diarrea.

Finalmente, **un análisis de comparación de costo** tomará múltiples programas y los comparará usando la misma unidad – permitiendo a los encargados de realizar políticas públicas preguntar: ¿Cuánto cuesta, por dólares, la reducción de la diarrea de cada estrategia?

See the paper on "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education" for more information.

OBJETIVOS, RESULTADOS Y MEDICIONES

Cuando se realiza una evaluación de programa, a menudo a los gobiernos y las ONGs se les piden destilar la misión de un programa a un puñado de *resultados* que, se entiende, se utilizarán para definir su éxito. Además de esta dificultad, cada resultado debe ser simplificado aún más a un *indicador* como la respuesta a una pregunta de la encuesta, o al resultado de una prueba.

Más que ser una labor de grandes proporciones, esto puede parecer imposible y la petición absurda. En el proceso, los evaluadores pueden parecer preocuparse sólo acerca de los datos y las estadísticas – no de la vida de las personas afectadas por el programa.

Para algunos objetivos, los indicadores correspondientes resultan ser naturales. Por ejemplo, si el objetivo de la distribución de tabletas de cloro es el de reducir las enfermedades transmitidas por el agua, el resultado relacionado puede ser una *reducción de la diarrea*. El indicador correspondiente, *incidencia de la diarrea*, podría venir de una pregunta en una encuesta en el hogar donde a los encuestados se les pregunta directamente, "¿Alguno de los miembros de su familia sufrió de diarrea en la semana pasada?"

Para otros objetivos, tales como "empoderar a la mujer", o "mejorar el civismo" los resultados no caen tan fácilmente en su lugar. Esto no significa que muchos objetivos son inmensurables. Por el contrario, se requiere más reflexión y creatividad para diseñar el indicador correspondiente. Para tener ejemplos de resultados difíciles de medir, vea el **artículo** adjunto.

¿QUÉ ES LA ALEATORIZACIÓN?

En el sentido más simple, la aleatorización es lo que sucede cuando se lanza una moneda, un dado, o cuando se hace una lotería, que determina qué es lo que pasa a continuación. Tal vez el resultado de esa moneda determina quién debe hacer alguna tarea; el dado determina quién recibe un monto de dinero; o la lotería determina quién participa en una actividad, o una encuesta. Cuando estas herramientas (la moneda, el dado o la lotería) se usan para tomar decisiones, se puede decir que el resultado se dejó en manos del azar, o que el resultado es **aleatorio**.

¿Por qué la gente deja que el azar determine su destino? Algunas veces, porque lo perciben como justo. Otras veces, porque la incertidumbre agrega un elemento de excitación. Los Estadísticos usan la aleatorización porque, cuando una cantidad suficiente de personas son *seleccionadas aleatoriamente* para participar en una encuesta, convenientemente, los atributos de esos individuos elegidos son *representativos* del grupo entero del que fueron elegidos. En otras palabras, lo que se descubra en ellos es probablemente cierto acerca del grupo más grande. Usar la lotería para obtener una muestra representativa es conocido como *muestreo aleatorio* o *selección aleatoria*.

Cuando dos grupos son seleccionados aleatoriamente de la misma población, *ambos representan* el grupo grande. No son sólo *estadísticamente equivalentes* al grupo grande; sino que también son estadísticamente equivalentes uno del otro. La misma lógica se lleva adelante si más de dos grupos son seleccionados aleatoriamente. Cuando dos o más grupos son seleccionados de esta forma, podemos decir que los individuos fueron *asignados aleatoriamente* a los grupos; esto se llama *asignación aleatoria* (asignación aleatoria es también el término apropiado cuando *todos* los individuos de un grupo grande son divididos aleatoriamente en diferentes grupos. Tal como antes, todos los grupos *representan* el grupo grande y son estadísticamente equivalentes el uno del otro). La *asignación aleatoria* es el elemento clave de la evaluación aleatoria.

Lo que sucede después en una evaluación aleatoria simple (con dos grupos) es que un grupo recibe el programa que está siendo evaluado y el otro no. Si estuviéramos por evaluar un programa de purificación de agua utilizando este método, asignaríamos aleatoriamente individuos a los dos grupos. Al inicio, los dos grupos serían estadísticamente equivalentes (y se espera que tengan trayectorias equivalentes hacía el futuro). Pero después introducimos algo que hace que sean diferentes; un grupo recibió el programa de purificación de agua y el otro no. Después de un tiempo, podríamos medir la salud relativa de los individuos en los dos grupos. Debido a que ellos son estadísticamente equivalentes al principio, las diferencias posteriores sólo pueden atribuirse a la entrega del servicio de purificación de agua.

El por qué se usa este método es un tema que será cubierto en la sección **¿Por qué Aleatorizar?.**

Las Evaluaciones Aleatorias tienen varios nombres:

- Pruebas de Evaluación Aleatorias
- Experimentos Sociales
- Estudios de Asignación Aleatoria
- Pruebas de Campo Aleatorias
- Experimentos Aleatorios Controlados

Las Evaluaciones Aleatorias son parte de un set más grande de evaluaciones llamadas **Evaluaciones de Impacto**. Las evaluaciones aleatorias a menudo se consideran el estándar de oro de las evaluaciones de impacto, porque siempre producen resultados más precisos.

Como todas las *evaluaciones de impacto*, el propósito principal de la aleatorización es la de determinar si un programa tiene impacto, y más específicamente, cuantificar *cuán grande* es el impacto. Las evaluaciones de impacto miden la efectividad del programa, típicamente comparando los resultados de aquellos (individuos, comunidades, escuelas, etc.) que participaron en el programa frente a aquellos que no lo hicieron. Hay varios métodos para hacer esto.

Lo que distingue las evaluaciones aleatorias de las que no lo son es que la participación (y no participación) es determinada *aleatoriamente* – antes de que el programa inicie. Esta *asignación aleatoria* es el método usado en las pruebas médicas para

determinar quién recibe un medicamento y quién recibe un placebo cuando se mide la efectividad (y efectos colaterales) de una nueva medicina. De la misma forma que en las pruebas médicas, aquellos en el programa que fueron *asignados aleatoriamente* al "grupo tratamiento" son elegibles para recibir el tratamiento (es decir, el programa); y son comparados con aquellos que aleatoriamente fueron asignados al "grupo control"- aquellos que no reciben el programa. Debido a que los miembros de los grupos (tratamiento y control) no difieren sistemáticamente desde el principio del experimento, cualquier diferencia subsecuente que surja entre ellos se atribuye al tratamiento más que a cualquier otro factor. Frente a los resultados de estudios no experimentales, los resultados de los estudios experimentales son:

- Menos sujeto a debates metodológicos
- Más fáciles de transmitir
- Más probable de ser convincentes a personas que financian programas y actores de políticas públicas.

Más allá de cuantificar los resultados causados por un programa, las evaluaciones aleatorias pueden cuantificar la incidencia de efectos secundarios no deseados (buenos o malos). Y al igual que otros métodos de evaluación de impacto, las evaluaciones aleatorias pueden dar una idea del por qué el programa falla o ha fallado en tener el impacto deseado.

1. Aleatorización en el contexto de "Evaluación"

Las evaluaciones aleatorias son un tipo de evaluación de impacto que usa una metodología específica para crear un grupo de comparación – en particular, la metodología de la asignación aleatoria. Las evaluaciones de impacto son evaluaciones de programas que se enfocan en medir los resultados finales de un programa. Hay muchos tipos de evaluaciones que pueden ser relevantes a los programas – más allá de medir la efectividad. (Vea ¿Qué es una Evaluación?)

2. Metodología de Aleatorización

Para entender mejor cómo funciona la metodología, vea Cómo realizar una evaluación aleatoria.

POR QUÉ

El propósito de las evaluaciones no es siempre claro para aquellos que vieron la realización de las encuestas, entraron la información, y que después entregaron reportes escritos que son rápidamente archivados para no ser nuevamente vistos. Lo único que se muestra a lo largo de todo el ejercicio es que el dinero, que pudo haber sido usado para expandir el programa, ahora ha desaparecido en esta evaluación y ya no está disponible. Esta historia es más común cuando las evaluaciones son impuestas por otros.

Si, por otro lado, aquellos responsables de tomar decisiones acerca de cómo diseñar el programa o aquellos que deciden qué programas implementar, tienen preguntas críticas, las evaluaciones pueden ayudarlos a encontrar las respuestas. Una evaluación es más útil cuando los encargados de un programa o los responsables de hacer políticas públicas están conduciendo la discusión acerca de qué debería ser evaluado. Se puede encontrar más información en la sección **¿Por qué Evaluar?**

Tal vez, una evaluación que hace las preguntas equivocadas es tan frustrante como una que hace las preguntas correctas pero produce respuestas no fiables. Montos significativos de dinero, tiempo, reflexión y esfuerzo se van en la búsqueda de encontrar las preguntas adecuadas. No es mucho pedir por respuestas precisas. En muchos casos, si se utiliza una metodología equivocada, incluso las técnicas estadísticas más elegantes no podrán corregir esos errores. Un diseño aleatorio puede ayudar a asegurar que las respuestas son fiables.

¿POR QUÉ EVALUAR?

El propósito de las evaluaciones no es siempre claro, en particular para aquellos que vieron la realización de las encuestas, entraron la información, y que después entregaron reportes escritos que son rápidamente archivados para no ser nuevamente vistos. Esto es más común cuando las evaluaciones son impuestas por otros.

Si, por otro lado, aquellos responsables de las operaciones del día a día de un programa tienen preguntas críticas, las evaluaciones pueden ayudar a encontrar las respuestas. Como ejemplo, la ONG responsable de la distribución de tabletas de cloro podría hablar con su equipo de trabajo local y escuchar historias de hogares que usan diligentemente las tabletas, y ocasionalmente ver mejoras en la salud. Pero cada vez que llueve fuerte, los hospitales se llenan de personas que sufren de diarrea. La ONG podría preguntarse, "si las personas están usando cloro para tratar el agua, ¿Por qué están enfermándose cuando llueve?. Aún cuando el agua pueda estar más contaminada, las tabletas de cloro son efectivas para matar bacterias." La ONG podría preguntarse si las pastillas de cloro son realmente efectivas para matar bacterias. ¿Estarán las personas utilizando la proporción adecuada? Tal vez nuestros empleados en terreno no nos están diciendo la verdad. Tal vez los beneficiaros no están usando las tabletas. Tal vez ni siquiera están recibiendo las tabletas. Y al confrontar estos hechos, los empleados en terreno se quejan de que durante las lluvias es difícil llegar a los hogares y distribuir tabletas. Los hogares, por otro lado, van a responder que ellos usan las tabletas durante las lluvias y que les ayudan bastante.

Hablar con individuos en distintos niveles de la organización así como con grupos de interés puede permitirnos descubrir muchas historias acerca de qué está pasando. Estas historias pueden ser la base de teorías. Pero explicaciones plausibles no son lo mismo que respuestas. Las evaluaciones incluyen el desarrollo de hipótesis acerca de qué está pasando, para después probar estas hipótesis.

¿POR QUÉ ALEATORIZAR?

¿Qué es impacto? En nuestro ejemplo del cloro, impacto es cuánto más saludables están las personas gracias al programa, de lo que estarían si no se hubiese aplicado el mismo. O más específicamente, en cuánto disminuyó la incidencia de diarrea de lo que lo hubiese hecho si no se hubiera aplicado el programa.

Obtener este número correctamente es más difícil de lo que parece. Es posible medir la incidencia de la diarrea en una población que recibe el programa, pero el "¿Qué hubiese pasado sin éste?" (denominado, el contrafactual) es imposible de medir directamente, sólo puede ser inferido.

Construyendo el grupo de Comparación

Las evaluaciones de impacto estiman la efectividad del programa usualmente al comparar los resultados de aquellos (individuos, comunidades, escuelas, etc.) que participaron en el programa frente a aquellos que no lo hicieron. El desafío clave en la evaluación de impacto es encontrar un grupo de personas que no participaran, pero con características lo suficientemente cercanas a la de los participantes, y en particular, a los participantes *si no hubiesen recibido el programa*. Medir los resultados en este grupo de comparación es lo más cercano que podemos estar de medir "cómo estarían los participantes si no hubiesen recibido el programa". Es por esto, que nuestra estimación del impacto es tan buena como nuestro grupo de comparación es equivalente.

Hay muchos métodos para crear grupos de comparación. Algunos métodos funcionan mejor que otros. Con todo lo demás igual, las evaluaciones aleatorias son las que funcionan mejor. Generan grupos de comparación *estadísticamente idénticos*, y por ende producen los resultados más precisos (sin sesgo). O dicho de otra forma: otros métodos, a menudo, producen resultados engañosos – resultados que llevarían a los responsables de la creación de políticas públicas a tomar las decisiones opuestas a lo que la verdad les hubiese mostrado.

Estos otro métodos no *siempre* nos dan la respuesta errónea, pero descansan sobre más supuestos. Cuando los supuestos se mantienen, las respuestas no tienen sesgo. Pero es normalmente imposible, y siempre difícil, asegurar que los supuestos son verdaderos. De hecho, es probable que la mayoría de los debates acerca de la validez de una evaluación giran en torno a los desacuerdos sobre la racionalidad de los supuestos.

Más allá de evitar debates acerca de los supuestos, las evaluaciones aleatorias producen resultados que son muy fáciles de explicar.

Aquí se muestra una tabla de comparación de los métodos de evaluación comúnmente usados.

¿QUIENES?

Cada evaluación aleatoria (EA) se hace posible a través de una asociación entre investigadores, organizaciones que ejecutan los programas a ser evaluados (como gobiernos o ONG), donantes, quienes financian los programas de investigación, centros de investigación, quienes emplean el personal asociados con cada evaluación y los sujetos de investigación que están de acuerdo en participar. Los programas sociales que evalúan las EA a menudo son diseñados para ser dirigidos a cierta población, por ejemplo, los pobres o los desamparados. Las poblaciones objetivo de estos programas también son los sujetos de investigación que participan en las EA.

Para una visión general de los principales actores que conducen EAs, haga clic **aquí**.

Para mayor información sobre las personas que participan en las EAs como sujetos de investigación, por favor haga clic **aquí**.

¿QUIÉN CONDUCE LAS EVALUACIONES ALEATORIZADAS?

J-PAL se fundó en 2003 como una red de **profesores afiliados** que conducen evaluaciones de impacto usando la metodología de evaluación aleatoria (EA), con el propósito de responder preguntas críticas relacionadas con el alivio de la pobreza. Los afiliados de J-PAL también conducen investigaciones no aleatorias, y muchas otras personas e instituciones conducen EAs. Para una breve historia del camino recorrido por las EA desde ensayos clínicos a experimentos agrícolas a programas sociales a alivio de la pobreza, haga clic **aquí**. Para una breve historia de J-PAL, haga clic **aquí**.

Desde la fundación de J-PAL, más de 200 organizaciones se han unido a un afiliado de J-PAL en alguna EA. Entre los actores claves para el alivio y desarrollo de la pobreza, el concepto de EA hoy es bastante conocido.

De las diez principales *fundaciones de los Estados Unidos,*[1] cuatro de las seis que trabajan en desarrollo internacional han trabajado con un afiliado de J-PAL en alguna EA. La Bill & Melinda Gates Foundation, la Ford Foundation, la William and Flora Hewlett Foundation, y la John D. and Catherine T. MacArthur Foundation[2] se incluyen entre estas fundaciones.

De las diez principales *organizaciones multilaterales*,[3] cuatro se han unido con un afiliado de J-PAL en alguna EA (el World Bank, el Asian Development Bank, Unicef, y el Banco Interamericano de Desarrollo), y seis de las diez han enviado personal a los cursos ejecutivos de J-PAL.

De las "Ocho Grandes" organizaciones de beneficencia,[4] Save the Children, Catholic Relief Services, CARE, y Oxfam se han unido con un afiliado de J-PAL en alguna EA. El International Rescue Committee se encuentra haciendo EAs por su cuenta. Y seis de las ocho han enviado personal a los cursos ejecutivos de J-PAL.

Gobiernos también se han unido a afiliados de J-PAL. Los principales socios de países donantes incluyen los Estados Unidos (USAID, MCC), Francia (Le Ministère de la Jeunesse et des Solidarités Actives), Suecia y el Reino Unido (DFID). Los socios de gobiernos de países en desarrollo han sido tanto a nivel nacional (Ej. Ministerio de la Educación de Kenia y la Secretaría de Descentralización del Gobierno de Sierra Leona) como a nivel sub-nacional (Ej. el Gobierno de Andhra Pradesh, el Pollution Control Board de Gujarat, y la policía de Rajasthan).

Se han establecido varios centros con el apoyo o bajo la dirección de los afiliados de J-PAL. Estos centros de investigación a menudo ejecutan las EA de los investigadores afiliados y emplean el personal relacionado con cada EA. Estos centros de investigación incluyen: Innovations for Poverty Action (IPA), Centre for Microfinance, Center for International Development's Micro-Development Initiative, Center of Evaluation for Global Action, Ideas42, y el Small Enterprise Finance Center.

Las empresas privadas también conducen evaluaciones aleatorias de programas sociales. Dos ejemplos de ello son Mathematica Policy Research y Abt Associates.

[1] Cuando se mide por donación.

[2] Las otras dos que trabajan en desarrollo internacional, pero que no se han unido con J-PAL son la Fundación W.K. Kellogg y la Fundación David and Lucile Packard. Las cuatro que hemos considerado que tienen un foco local en Estados Unidos son Getty Trust, Robert Wood Johnson Foundation, Lilly Endowment Inc., y Andrew W. Mellon Foundation.

[3] Cuando se mide por asistencia oficial al desarrollo otorgada, incluyendo Banco Mundial, Grupo Banco Africano de Desarrollo, The Global Fund, Banco Asiático de Desarrollo, Fondo Monetario Internacional, Unicef, UNRWA, Banco Interamericano de Desarrollo, Programa de las Naciones Unidas para el Desarrollo, y World Food Program.

[4] Cuando se mide por presupuesto anual. Estas son World Vision, Save the Children, Catholic Relief Services, CARE, Medecins Sans Frontieres, Oxfam, International Rescue Committee, y Mercy Corps.

¿QUIÉN PARTICIPA EN LAS EVALUACIONES ALEATORIAS?

La pregunta de quién participa en una evaluación aleatoria involucra a algunos de los asuntos más delicados confrontados por un evaluador. Al responder esta pregunta, el evaluador debe considerar qué es ético y justo. Sería poco ético, por ejemplo, privar a un hogar de una solución de tratamiento de agua por motivos de un experimento cuando de otra forma éste sí hubiera tenido acceso al servicio.

1. Asuntos Éticos

Entonces, ¿Cómo puede un evaluador conducir un experimento y también llevar estándares de ética y justicia?

Las evaluaciones aleatorias pueden ser apropiadas en situaciones en las cuales existen recursos restringidos. Típicamente, una organización no tiene suficiente presupuesto como para aplicar un programa a toda una comunidad o distrito o país. Debido a restricciones presupuestarias, la organización debe decidir quién recibe el programa y quién no. Incluso si determinan un subgrupo de personas que necesitan más el programa, o que se verían más beneficiados, probablemente no sean capaces de cubrir a todos aquellos pertenecientes a estos subgrupos. Esto brinda al evaluador la oportunidad de llevar a cabo una evaluación aleatoria. Un evaluador puede decidir aleatoriamente cómo asignar los recursos escasos dentro del sub grupo objetivo.

Un evaluador no sólo debe asegurarse de que el experimento sea ético, sino además que sea justo. Al asignar a los participantes a los grupos de control o de tratamiento, un evaluador debería asegurarse que todos tengan iguales probabilidades de estar en el grupo experimental y recibir el tratamiento. Dentro de los métodos justos para seleccionar participantes están las loterías, las introducciones graduales de programas, y la rotación de participantes dentro del programa para asegurar que todos reciban los beneficios. El proceso de selección también debiera ser transparente y debe parecerle justo a la comunidad.

Típicamente los evaluadores se enfrentan con el problema de asignar programas que son claramente beneficiosos, como la desparasitación, o soluciones de tratamiento de aguas. En otras palabras, el dilema ético surge cuando se crea un grupo de individuos a los cuales se les negará el programa. Algunas veces, sin embargo, los beneficios no han sido probados, lo que significa que es posible que el programa pueda potencialmente empeorar la situación de los individuos. Por ejemplo, las compañías de medicamentos suelen tener este problema cuando prueban nuevos tratamientos en sus pacientes. En este caso, un evaluador debe

poner mucha energía en asegurar que los pacientes en el grupo de tratamiento no serán dañados. De existir un riesgo potencial para los participantes, entonces todos los involucrados deben ser informados sobre los riesgos, y sus consentimientos son necesarios para participar. Incluso si no pareciera haber riesgos, todo experimento debiera requerir la información y el consentimiento de todos los participantes (tanto en grupos de comparación como de tratamiento). Diversas naciones y organizaciones han desarrollado protocolos para los seres humanos, y éstos deben ser respetados. (Ver más abajo)

2. Sujetos de Investigación y la Comisión de Revisión Internacional (Institutional Review Board)

Una Comisión de Revisión Internacional (IRB), también conocida como comité de ética independiente o comisión de revisión de sujetos humanos, es un grupo que ha sido designado formalmente por una institución (como una universidad u organización sin fines de lucro) con el propósito de aprobar, monitorear y revisar la investigación que involucra a humanos como participantes. El objetivo de una IRB es garantizar, tanto antes de la implementación y en revisiones periódicas, que se toman los acciones correspondientes para proteger los derechos y bienestar de los humanos que participan como sujetos en una investigación.

Debido a que los estudios de J-PAL involucran participantes humanos, los asociados de J-PAL y su personal garantizan que sus estudios cumplan con las pautas de los métodos éticos de investigación, los cuales incluyen:

- Recepción de las aprobaciones de la Comisión de Revisión Internacional (IRB) para cada estudio antes de que éste comience,
- Todo personal de estudio realiza curso de capacitación de la IRB,
- Seguimiento del protocolo y pautas de investigación aprobados por la IRB a lo largo del estudio.

¿CUÁNDO?

Para una pequeña reseña sobre la historia de las evaluaciones aleatorias, ver **"¿Cuándo comenzaron las evaluaciones aleatorias?"**

Para leer cuándo son apropiadas las evaluaciones aleatorias, ver : **"¿Cuándo conducir una evaluación aleatoria?"** o **"¿Cuándo (no) es apropiada la aleatorización?"**

¿CUÁNDO COMENZARON LAS EVALUACIONES ALEATORIAS?

1. Ensayos Clínicos

El concepto de grupo experimental y de control fue introducido en 1747 por James Lind cuando demostró los beneficios de los frutos cítricos para prevenir el escorbuto a través de un experimento científico.[1] Por los resultados de su trabajo, Lind es considerado como el padre de los ensayos clínicos. El método de asignación aleatoria a grupos de control y tratamiento, sin embargo, no se desarrolló sino hasta la década de 1920.

2. Experimentos Agrícolas

La aleatorización se introdujo en la experimentación científica en la década de 1920 cuando Neyman y Fisher condujeron las primeras pruebas aleatorias en experimentos agrícolas. El experimento de campo de Fisher culminó con su libro emblemático, El Diseño de los Experimentos, que impulsó en gran medida el crecimiento de las evaluaciones aleatorias.[2]

3. Programas Sociales

Las pruebas aleatorias fueron introducidas para realizar experimentos sociales patrocinados por los gobiernos entre 1960 y 1990. En vez de tratarse de experimentos de pequeña escala en animales y plantas, estos experimentos eran de escala significativamente

mayor, y enfocados en personas como objeto de interés. La idea de conducir experimentos para programas sociales creció a partir de un debate en la década de los 60 sobre los méritos del sistema de bienestar social. El modelo de experimentación social fue aplicado más tarde tanto en Europa como en los Estados Unidos para evaluar otros programas tales como diseños de esquemas de precios de la electricidad, programas de desempleo, y de subsidios de vivienda. Desde entonces, los experimentos sociales son usados en diversas disciplinas y en una variedad de contextos alrededor del mundo para guiar las decisiones de políticas públicas.[3]

El Abdul Latif Jameel Poverty Action Lab (J-PAL) fue fundado en Junio de 2003 como una red de profesores afiliados de todo el mundo, a quienes los une el uso de **evaluaciones aleatorias** para responder preguntas esenciales para la reducción de la pobreza

[1]Thomas, Duncan P. Sailors, Scurvy y Science. Journal of the Royal Society of Medicine. 90 (1997).

[2] Levitt, Steven D. y List, John A.,Field Experiments in Economics: The Past, the Present, and the Future (Septiembre de 2008). NBER Working Paper No. W14356. Disponible en SSRN: http://ssrn.com/abstract=1271388

[3] ibid

¿CUÁNDO CONDUCIR UNA EVALUACIÓN?

El valor agregado de evaluar una política pública rigurosamente depende del momento en el ciclo de vida del programa en el que se lleva a cabo dicha evaluación. La evaluación no debiera ser muy temprano: cuando el programa aún está tomando forma y sus aristas están siendo perfeccionadas. Tampoco debiera ser muy tarde: después de que los fondos han sido asignados y que el programa se ha desplegado, de manera que no hay ya espacio para un grupo de control.

El tiempo ideal es durante la fase piloto de un programa, o antes de aumentar la escala de éste. Durante estas etapas surgen preguntas importantes que a un evaluador le gustaría poder contestar: ¿Qué tan efectivo es el programa?, ¿Es efectivo en distintas poblaciones?, ¿Hay algunos factores que funcionan mejor que otros?, y ¿pueden "los otros" ser mejorados?, ¿Es el programa efectivo cuando se aplica a una población más grande?

Durante la fase piloto, los efectos de un programa sobre una población determinada son desconocidos. El programa incluso podría ser nuevo, o podría ser uno antiguo que se aplica a una nueva población. En ambos casos, los jefes del programa y los diseñadores de políticas públicas quisieran comprender mejor la efectividad del programa y cómo puede ser mejorado. Casi por definición, el programa piloto se aplica sólo a una porción de la población objetivo, lo que hace posible la realización de un experimento aleatorio. Luego de la fase piloto, si el programa ha resultado ser efectivo, conduciendo a un mayor apoyo y a una mayor asignación de recursos, entonces el programa puede ser replicado o expandido a todo el resto de la población objetivo.

Un ejemplo de una evaluación aplicada en el momento adecuado es el de PROGRESA, un programa de transferencias monetarias condicionadas, aplicado en México en 1997. El programa daba subsidios en efectivo a las madres siempre y cuando éstas se aseguraran de que sus hijos fueran regularmente a la escuela y recibieran vacunas programadas. El partido político que había estado en el poder por 68 años, Partido Revolucionario Institucional (PRI), estaba enfrentando una derrota inminente en las elecciones venideras. Un resultado probable de la derrota electoral era el desmantelamiento de programas como PROGRESA. Para poder defender este programa, el PRI tuvo que demostrar claramente la efectividad de la política pública para mejorar la salud y la educación de los niños.

PROGRESA fue introducido primero como un programa piloto en áreas rurales en 7 estados. De las 506 comunidades escogidas por el gobierno mexicano para el piloto, 320 fueron aleatoriamente asignadas al grupo de tratamiento y 186 al de control. Al comparar ambos grupos al cabo de un año, se encontró que los niveles de salud y educación de los niños tratados eran mejores. Como era de esperar, la popularidad del programa se expandió desde sus defensores iniciales y beneficiarios directos hasta la totalidad de la nación.

Luego de la esperada derrota del PRI en las elecciones de 2000, el nuevo partido político (PAN) tomó el poder y heredó un programa de gran popularidad. En vez de desmantelar PROGRESA, el PAN le cambió el nombre a OPORTUNIDADES, y lo expandió a toda la nación.

El programa se replicó rápidamente en otros países, como Nicaragua, Ecuador y Honduras. Además, siguiendo la pauta de México, estos nuevos países condujeron estudios piloto para evaluar el impacto de otros programas como PROGRESA antes de replicarlos a gran escala.

¿CUÁNDO (NO) ES APROPIADA LA ALEATORIZACIÓN?

Las Evaluaciones Aleatorias pueden no ser apropiadas:

1. Cuando se evalúan políticas macro.

Ningún evaluador tiene el poder político para conducir una evaluación aleatoria de distintas políticas monetarias. No se puede asignar aleatoriamente un tipo de cambio flotante al Japón y otras naciones, y un tipo de cambio fijo a los Estados Unidos y otro grupo de naciones.

2. Cuando es poco ético o políticamente imposible negarle el programa al grupo de control.

No sería ético negar un medicamento con beneficios ya comprobados a un grupo de pacientes si es que se tienen los recursos para darlo.

3. Si el programa cambia durante el curso del experimento.

Si a la mitad de un experimento el programa cambia de ofrecer solución de tratamiento de aguas a ofrecer tratamiento de aguas y una letrina, se vuelve difícil interpretar qué parte del programa causó los resultados observados.

4. Cuando el programa en su fase experimental es significativamente diferente a como se espera que sea el programa en condiciones normales.

Durante un experimento es más probable que los participantes usen la solución de tratamiento de agua si se les da incentivos. En condiciones normales, sin incentivos será menor el número de personas que realmente usen la solución de tratamiento de agua, incluso si ya la tienen y saben como usarla.

A modo de advertencia, este tipo de evaluación puede servir como Prueba de Concepto. Sencillamente trataría de responder la pregunta "¿puede este programa o política ser efectivo?". No se esperaría que arroje resultados generalizados.

5. Cuando una EA consume demasiado tiempo o es muy costosa, por lo tanto no es costo-efectiva.

Por ejemplo, debido a una política de gobierno, una organización puede no tener suficiente tiempo para hacer un programa piloto y evaluarlo antes de su implementación.

6. Cuando la atrición o el efecto de las externalidades son demasiado difíciles de controlar y esto daña la integridad del experimento.

Una organización puede decidir evaluar el impacto de un medicamento para eliminar parásitos sobre la asistencia a clases en un colegio particular. Debido a que los medicamentos de desparasitación tienen un efecto de externalidad (la salud de un estudiante afecta en la salud de otros), será difícil medir adecuadamente el impacto del medicamento. En este caso, una solución podría ser la aplicación del programa a nivel de escuela y no a nivel de alumno.

7. Cuando el tamaño de la muestra es muy pequeño.

Si hay demasiado pocos sujetos participando en el programa piloto, incluso si el programa fue exitoso, no hay suficientes observaciones como para estadísticamente detectar un impacto.

CÓMO CONDUCIR UNA EVALUACIÓN ALEATORIA

Algunos se refieren a las evaluaciones aleatorias como el *estándar de oro* de las evaluaciones de impacto, porque son irrevocablemente las más rigurosas – queriendo decir que son las que requieren menos supuestos, o menos saltos de fe, cuando se sacan conclusiones de los resultados. Ser la más rigurosa no significa sin embargo ser la que requiere más trabajo o costo. De hecho, asignar a los individuos a los grupos de forma aleatoria para asegurar que sean equivalentes al principio (ver **¿Qué es Aleatorizar?** y **¿Por qué Aleatorizar**?) puede reducir la cantidad de trabajo estadístico para sintetizar un grupo de comparación equivalente más adelante en la fase de análisis.

Existen algunos desafíos al conducir una evaluación aleatoria: convencer a los ejecutores del programa de aleatorizar, pensar sobre el diseño más apropiado para el experimento, asegurar que la integridad del diseño de la evaluación (la asignación aleatoria) se mantenga. Pero la mayor parte del trabajo y costo viene de asegurarse una muestra de tamaño suficiente como para detectar un impacto (un requisito también para las evaluaciones no aleatorias) y descubrir qué hace funcionar o fallar al programa.

PLANEANDO UNA EVALUACIÓN

Al planear una evaluación es importante identificar las preguntas claves que la organización quiere responder. De éstas, podemos determinar cuántas pueden ser respondidas revisando las evaluaciones de impacto previas o de un sistema mejorado de **evaluación de procesos**. Asumiendo que no podamos responder todas nuestras preguntas, debemos entontes escoger algunas que tengan mayor prioridad, que serán el principal objetivo de nuestra evaluación de impacto. Finalmente debemos elaborar planes para responder la mayor cantidad posible de estas preguntas, teniendo siempre en cuenta que unos pocos estudios de impacto de alta calidad son más valiosos que muchos estudios de baja calidad.

El primer paso en una evaluación es revisar las metas del programa y cómo esperamos alcanzarlas. Un marco lógico o un modelo de teoría de cambios son útiles en este proceso (ver **Evaluación Teórica del Programa**). Al evaluar el propósito y estrategia de un programa, debemos identificar los resultados clave, los caminos esperados para lograr aquellos resultados, y algunos hitos que nos indiquen que vamos por buen camino. Como es de esperar en una evaluación, estos resultados e hitos necesitarán ser medidos, y por lo tanto transformados en indicadores y, finalmente, en información (ver **Objetivos, Resultados y Mediciones**).

Sólo después de tener una buena noción de las vías y ámbitos de influencia, y de tener planificado cómo medir nuestros progresos, podemos pensar en **el diseño de la evaluación**.

CÓMO DISEÑAR UNA EVALUACIÓN

El diseño de una evaluación requiere una cantidad considerable de pensamiento. Primero viene la parte conceptual: ¿Qué planeamos aprender de esta evaluación?, ¿Cuáles son las preguntas relevantes? , ¿Qué resultados se esperan?, ¿Cómo se pueden medir?

A continuación, vienen las preguntas del diseño:

¿Cuál es el *nivel* o la **unidad de aleatorización adecuada**?

¿Cuál es el **método de aleatorización adecuado**?

Además de las restricciones políticas, administrativas y éticas, ¿qué aspectos técnicos pueden comprometer la integridad de nuestro estudio, y como podemos mitigar estas **amenazas en el diseño**?

¿Cómo haremos para **implementar la aleatorización**?

¿Cuál es el **tamaño de muestra** necesario para responder nuestras preguntas? (¿cuánta gente debemos incluir en nuestro estudio, no sólo como participantes sino también como contestadores de encuestas?)

1. Unidad de Aleatorización

Al diseñar nuestra evaluación debemos decidir a qué nivel haremos la aleatorización: ¿cuál será la unidad sujeta a asignación aleatoria? ¿Serán individuos o grupos de individuos, tales como hogares, pueblos, distritos, escuelas, clínicas, grupos de iglesia, empresas y asociaciones de crédito? (Cuando la unidad de aleatorización es un grupo de individuos – incluso cuando nos interesa la medición de resultados individuales – nos referimos a *evaluación aleatoria por clúster*). Por ejemplo, si logramos dar píldoras de cloro a mil hogares para tratar aguas contaminadas (de una muestra de, digamos, diez mil hogares que sacan agua de la misma fuente contaminada), ¿asignaríamos aleatoriamente a los *hogares* que serán tratados, dejando al resto en el grupo de control? Esto significaría que algunos hogares recibirían pastillas de cloro, mientras que algunos de sus vecinos más cercanos se quedarían sin este beneficio. ¿Es esto factible? ¿Ético?

Para este tipo de programa, probablemente tampoco sería posible hacer la asignación a menor nivel, por ejemplo a nivel individual. Implicaría que algunos niños dentro de un hogar reciban la píldora de cloro mientras que sus hermanos no. Si todos los miembros de un hogar beben del mismo tanque tratado de agua, la asignación aleatoria individual sería físicamente imposible, aun sin tomar en cuenta las consideraciones éticas.

Tal vez una medida apropiada de asignación aleatoria es la comunidad, según la cual algunas comunidades reciben cloro, otras no, pero dentro de una comunidad "en tratamiento" todos los hogares (lo que implica a todos los vecinos) son elegibles para recibir la píldora de cloro. Hay muchos aspectos a considerar cuando se determina el nivel apropiado de aleatorización, de los cuales la ética y la factibilidad son sólo dos. Siete aspectos son mencionados a continuación.

- **¿Qué unidad de tratamiento es la meta del programa?**
- **¿Cuál es la unidad de análisis?**
- **¿Es el diseño de la evaluación justo?**
- **¿Es la evaluación aleatoria políticamente factible?**
- **¿Es la evaluación aleatoria logísticamente factible?**
- **¿Qué efectos de externalidad u otros efectos deben ser tomados en cuenta?**
- **¿Qué tamaño de muestra y poder necesitamos para detectar los efectos del programa?**

1. *¿Qué unidad de tratamiento es la meta del programa?*: Si las tabletas de cloro se disuelven en tanques de agua que, en nuestra región, todos los hogares suelen poseer, entonces es posible que algunos hogares reciban las tabletas y otro no. En este caso, la unidad de asignación aleatoria sería a nivel de hogar. Sin embargo, si el tanque de agua suele estar ubicado fuera de las casas y usado por un grupo de hogares, sería imposible asignar algunos hogares de este grupo al grupo de control--todos beben la misma agua (tratada) que beben los hogares en el grupo de tratamiento. Entonces, la unidad más natural de asignación sería aquel "grupo de hogares" que comparten un tanque de agua.

2. *¿Cuál es la unidad de análisis?:* Si la evaluación tiene que ver con los efectos a nivel de la comunidad, entonces el nivel más natural de asignación es el comunitario. Por ejemplo, supongamos que medimos los resultados en cantidad de "hospitalizaciones" debido a la diarrea, y esto es más económico de medir usando los registros administrativos de las clínicas comunitarias que, además, son anónimos. No podríamos distinguir si las personas que se hospitalizaron pertenecían a los hogares en el grupo de tratamiento o de control. Sin embargo, si toda la comunidad está en el grupo de tratamiento, podríamos comparar los registros de las clínicas en comunidades tratadas contra las clínicas en comunidades no tratadas.

3. *Justicia en el diseño de la evaluación:* El programa debe ser percibido como uno justo. Si se me han negado las píldoras de cloro, pero mis vecinos más cercanos las reciben, estaré enojado con mis vecinos, estaré enojado con la ONG, y estaré menos dispuesto a rellenar cualquier cuestionario sobre el uso de cloro cuando los encuestadores vayan a mi casa a pedírmelo. Y a la ONG no estará contenta de hacer enojar a los miembros de su comunidad. Por otro lado, si nadie en mi comunidad salió beneficiado, pero la comunidad vecina sí, puede que nunca sepa nada al respecto y por lo tanto no tenga quejas, o puede que piense que fue una decisión tomada a nivel de pueblo, y que la mía decidió no invertir en pastillas de cloro. Por supuesto, las personas también podrían enojarse con una asignación a nivel comunitario. Podríamos tratar de expandir la unidad de asignación aleatoria, o pensar en otras estrategias para mitigar el descontento de la gente que no salió beneficiada. El hecho de que no todos son favorecidos puede ser injusto (ver **asuntos éticos**). Pero dado que no podemos ayudar a todos (usualmente debido a restricciones de capacidad), y nuestro deseo de mejorar y evaluar, podemos repartir los recursos de una forma que nos ayude a crear un grupo de control y que al mismo tiempo sea visto como justo por las personas que estamos tratando de ayudar.

4. *Factibilidad Política:* Puede que no sea factible hacer una asignación aleatoria a nivel de hogar. Por ejemplo, la comunidad puede exigir que toda persona en necesidad debe recibir asistencia, lo que hace imposible escoger aleatoriamente los hogares a los cuales darles las píldoras de cloro. En algunos casos, el líder solicita que todos los miembros de su comunidad reciban asistencia. O puede que se sienta más tranquilo si la mitad obtiene el beneficio al azar (con absoluta certeza, en el caso de asignación individual), que si afronta el riesgo de que nadie en su comunidad sea tratado (en el caso de que la asignación sea comunitaria y su pueblo no salga escogido). En algunos casos, el líder puede colaborar con el estudio; en otros, no.

5. *Factibilidad logística:* A veces es logísticamente imposible asegurarnos de que algunos hogares permanezcan en el grupo de control. Por ejemplo, si la entrega del cloro requiere que un distribuidor en cada pueblo monte un puesto donde los vecinos pueden ir a buscar sus píldoras, puede ser ineficiente pedirle que no considere a los hogares en el grupo de control. Esto puede añadir burocracia, perder tiempo y distorsionar la idea original del programa. Incluso si el distribuidor pudiera discriminar fácilmente, los hogares que reciben píldoras podrían compartirlas con sus vecinos que no fueron beneficiados. Entonces, el grupo de control se vería también impactado por el programa y no serviría como grupo de comparación. (Recordemos que el grupo de control supuestamente representa cómo sería la vida sin el programa) (ver ¿Qué es una **evaluación de impacto**?). En este caso, haría sentido asignar a nivel de pueblo, y sencillamente contratar distribuidores para que visiten los pueblos tratados y no los de control.

6. *Controlar las externalidades y otros efectos*: Incluso si es factible asignar a nivel de hogar –dar cloro en algunas casas y en otras no – puede no ser posible contener el impacto dentro del grupo de tratamiento. Si lo hogares del grupo de control son afectados por el hecho de que se lleve a cabo el programa – si se benefician por que hay menos enfermos (efecto de externalidad), o beben el agua de los hogares en tratamiento (no cumplen con la asignación aleatoria y se pasan al grupo de tratamiento), pues ya no son un buen grupo de comparación. (ver ¿Qué es una **evaluación de impacto**?) (para más detalles sobre efectos de externalidad o efecto control-tratado, ver **Amenazas al diseño**.)

7. *Tamaño de muestra y poder:* La habilidad de detectar efectos reales depende del tamaño de muestra. Cuanto mayor es el número de personas escogidas a partir de una gran población, estadísticamente, mejor representan a la a aquella población (ver **Selección y tamaño de la muestra**). Por ejemplo, si encuestamos a dos mil hogares, y aleatorizamos a nivel de hogar (mil hogares reciben tratamiento, mil hogares son el control), efectivamente tendremos un tamaño de muestra de dos mil hogares. Pero si aleatorizamos a nivel de pueblo, y cada pueblo tiene cien hogares, entonces tendremos 5 pueblos en el grupo de tratamiento y 5 en el grupo de control. En este caso, estaríamos midiendo los casos de diarrea e nivel de hogar, pero debido a que aleatorizamos a nivel de pueblo, puede ser que tengamos una muestra efectiva de 10 hogares (aunque hayamos encuestado a dos mil hogares!). En realidad, el tamaño efectivo de la muestra puede ser cualquiera entre diez y dos mil, dependiendo de qué tan parecidos sean los habitantes de un pueblo entre sí (Ver: tamaño de muestra). Con un tamaño de muestra efectivo de 10, no seríamos capaces de detectar efectos reales. Esto puede influenciar nuestra decisión con respecto de qué nivel de asignación usar.

Existen muchas consideraciones al determinar el nivel apropiado de asignación aleatoria. Los evaluadores no pueden simplemente sentarse frente a un computador, presionar un botón, producir una lista e imponer un diseño de evaluación para una organización que está a miles de kilómetros de distancia. Los evaluadores deben tener una comprensión profunda y completa de la organización responsable, del programa, y del contexto y del trabajo del equipo para determinar el nivel de asignación apropiado para cada circunstancia en particular.

2. Diferentes Métodos de Aleatorización

Si mi organización puede dar mil píldoras de cloro cada día, entonces puedo beneficiar a mil personas de un grupo de dos mil postulantes cada día, y puedo escoger beneficiar siempre a los mismos. Alternativamente, puedo ir rotando día por medio para que cada hogar pueda beber agua limpia día por medio. Puede ser que la última opción no me parezca razonable. Si todos beben agua sucia la mitad de los días, no esperaría ningún efecto sobre nadie. Entonces, puedo escoger a la mitad de los postulantes que recibirán la píldora de cloro perpetuamente. Para aleatorizar, puedo realizar una *lotería* simple para elegir los mil hogares que recibirán las píldoras: escribiré los nombres de las dos mil personas en pequeños trozos de papel, pondré estos pedazos en una caja, agitaré la caja, cerraré mis ojos y sacaré mil papeles. Intuitivamente, esto es lo que se conoce como diseño usando lotería.

Alternativamente, si quiero rotar los hogares que reciben el tratamiento cada año en vez de cada día, y asignar aleatoriamente el orden en el cual serán tratados, entonces en un año algunos hogares serán considerados dentro del grupo de tratamiento, y en el próximo serán parte del grupo de control. Si voy a medir los resultados al final de cada año, esto sería un diseño por *rotación*.

Digamos que este año puedo entregar quinientas píldoras de cloro cada día, pero para el próximo año espero poder entregar mil diarias, y el año siguiente dos mil diarias. Podría escoger aleatoriamente quinientos hogares para ser tratados el primer año, añadir otros quinientos que se sumen en el segundo año, y añadir a los mil hogares restantes el tercer año. Esto es lo que llamaríamos diseño *escalonado*.

Existen siete posibles modos de diseño de aleatorización –los diseños usando lotería, diseño escalonado, diseño por rotación, diseño por estímulos a participación, diseño con tratamientos con diferente intensidad, y la asignación aleatoria en dos etapas. Estos diseños no son necesariamente mutualmente excluyentes. Sus ventajas y desventajas vienen resumidas en la siguiente tabla.

3. Amenazas al diseño

(a) Externalidades

Un efecto de externalidad ocurre cuando un programa, diseñado para ayudar a un grupo objetivo, afecta de modo no intencional al grupo de control (ya sea positiva o negativamente). El grupo de control debería representar el resultado si el programa no hubiera sido implementado (ver **contrafactual**). Si este grupo de comparación ha sido alterado por el programa, su rol de emulador del contrafactual se ve comprometido, y la medición del impacto puede estar sesgada. Existen maneras de mitigar los efectos de externalidad, por ejemplo, cambiando el nivel de aleatorización.

Por ejemplo, beber aguas contaminadas puede provocar enfermedades. Pero jugar con los niños del vecindario que están enfermos también las provoca. Si estoy en el grupo de control y el programa beneficia a mis vecinos, esos vecinos ya no estarán enfermos, lo que reduce mi posibilidad de enfermarme. Entonces, incluso cuando estoy en el grupo de control, el tratamiento a mis vecinos me afecta indirectamente. Ya no soy un buen grupo de comparación. Esto es conocido como el efecto de externalidad, en particular, se trata de una externalidad positiva. Para mitigar esto, podríamos aleatorizar a nivel de comunidad. Hacer esto significaría que si nuestra comunidad fue asignada al grupo de control, yo y mis vecinos tendremos el mismo estatus. Tengo menos probabilidades de jugar con niños de otros pueblos, y por lo tanto tengo menos probabilidades de ser impactado indirectamente por el tratamiento. O, si nuestra comunidad fuera asignada al grupo de tratamiento, no podríamos impactar positivamente a los demás.

(Por supuesto, podría interesarnos conocer cómo ocurren estas externalidades, y hacer diseños acorde con esto. Ver **métodos de aleatorización**).

b) Efecto Control-Tratado

Otra posibilidad es que mi hogar haya sido asignado al grupo de control, pero mi vecino esté en el grupo de tratamiento, y por tanto mi madre sabe que su agua está limpia y me manda a su casa a beberla. De este modo, me infiltro en el grupo de tratamiento, aun cuando yo fui asignado al grupo de control. Cuando las personas deliberadamente desafían su designación de tratamiento (consciente o inconscientemente) los resultados son alterados, y se considera un efecto control-tratado. Al igual que con las externalidades, al cruzarme, yo ya no represento un buen grupo de comparación – ya que he sido afectado por la presencia del programa. Al igual que antes, cambiar el nivel de la aleatorización podría mitigar los efectos control-tratado.

4. Mecánica de la Aleatorización

Una vez que la unidad y el método de aleatorización se hayan determinado, es tiempo de asignar aleatoriamente a los individuos, hogares, comunidades o cualquier otra unidad, al grupo de tratamiento o de control.

a) Lotería simple

Generalmente para comenzar, necesitamos una lista de nombres (de individuos, jefes de hogares, o pueblos). Después, hay varias maneras de proseguir. Podríamos escribir todos los nombres en un pedacito de papel, ponerlo en una canasta, agitar la canasta, cerrar nuestros ojos y sacar mil pedacitos de papel. Esos conformarían nuestro grupo de tratamiento y el resto podría ser el grupo de control (o viceversa). Podríamos hacer esto como parte de una lotería pública. Similarmente, podríamos ir leyendo la lista, y uno a uno, con la ayuda de una moneda, determinar su estatus de tratamiento. Sin embargo, no siempre dividimos a la población del estudio exactamente a la mitad. Por ejemplo, quizás quisiéramos incluir 30 por ciento en el grupo de tratamiento y 70 en el grupo de control. O si tuviéramos un método escalonado en tres periodos, podríamos tratar de dividir la población en tres grupos. También muy común, es tratar de testear múltiples tratamientos al mismo tiempo – también requiriendo varios grupos. En estos diseños de evaluaciones más sofisticados, lanzar una moneda no va a ser suficiente.

Típicamente, se escribe un programa de computadora para aleatoriamente asignar nombres a los grupos.

b) Aleatorización instantánea

Algunas veces no tenemos una lista de nombres de antemano. Por ejemplo, si individuos entra en una clínica con síntomas de malaria, la decisión de administrar el tratamiento estándar de la Organización Mundial de la Salud "DOTS" o una alternativa mejorada, debe hacerse en el momento. El tratamiento podría ser determinado por la enfermera en la clínica tirando una moneda. Pero podemos estar preocupados de que la enfermera haga caso omiso de la asignación al azar si ella tiene una opinión de cuál tratamiento es mejor y qué pacientes son más "dignos" que otros. Las alternativas podrían incluir la aleatorización computarizada o por teléfono celular.

c) Aleatorización estratificada

Con frecuencia, la población objetivo se divide en subgrupos antes de aleatorizar. Por ejemplo, un grupo de individuos se pueden dividir en grupos más pequeños por razón de sexo, origen étnico o edad. O pueblos se pueden dividir en regiones geográficas. Esta división en subgrupos antes de la aleatorización se llama estratificación. Después, la aleatorización toma lugar dentro de cada uno de los estratos (subgrupos). Esto se hace para garantizar que los grupos de tratamiento y de control tienen proporciones equilibradas de tratamiento y control dentro de cada grupo. Es posible que con una pequeña muestra, nos encontramos con que, sin estratificar, hayan más mujeres en nuestro grupo de tratamiento que hombres. El propósito principal de la estratificación es estadística y se relaciona al tamaño de la muestra. La decisión de estratificar no influye en el tema de sesgo.

5. Selección y tamaño de la muestra

Un experimento debe ser lo suficientemente sensible para detectar diferencias de resultados entre el grupo de tratamiento y el de comparación. La sensibilidad de un diseño se mide por el poder estadístico, que, entre otros factores, depende del tamaño de la muestra - es decir, el número de unidades asignados aleatoriamente y el número de unidades encuestadas.

Una vez más, tomemos el ejemplo de las enfermedades transmitidas por el agua en una comunidad. Supongamos que hemos elegido distribuir pastillas de cloro a los hogares para comprobar su impacto en la incidencia de la diarrea. Pero también supongamos que sólo tenemos un presupuesto muy limitado para nuestra fase de prueba, y lo que nos gustaría es minimizar el número de hogares que se incluyen en la encuesta, pero sin dejar de asegurarnos que podemos saber si cualquier cambio en la incidencia se debe a las tabletas de cloro y no por casualidad. ¿Cuántos hogares deben recibir las tabletas y cuántos deben ser encuestados? ¿Son cinco hogares suficientes? ¿100? ¿200? ¿Cuántos hogares deben estar en el grupo de control? Las pruebas de poder estadística nos ayudan a responder estas preguntas.

Para obtener más información sobre cómo calcular el tamaño de muestra, vea:

Duflo, Esther, Glennerster, Rachel, and Kremer, Michael, "Using Randomization in Development Economics Research: A Toolkit" (2006). MIT Department of Economics Working Paper No. 06-36.

Bloom, H.S. (1995): "Minimum Detectable Effects: A simple way to report the statistical power of experimental designs," Evaluation Review 19, 547-56.

COMO IMPLEMENTAR UNA EVALUACIÓN

Una vez que se ha finalizado el diseño de la evaluación, el evaluador debe seguir participando en el monitoreo de la recolección de datos, así como en la implementación de la intervención que se está evaluando. Si los encuestados desaparecen durante la fase de recolección de datos, los resultados son susceptibles a un **sesgo de atrición**, comprometiendo su validez. La atrición se cubre en esta sección. Otras amenazas durante la fase de recopilación de datos como: instrumentos de medición pobres, sesgos de reporte, etc. son igualmente importantes, pero no se cubren aquí. Para aprender sobre las mejores prácticas en la recolección de datos, vea:

Deaton, A. (1997): The Analysis of Household Surveys. World Bank, International Bank for Reconstruction and Development

En la ejecución de la intervención, la integridad de la aleatorización debe permanecer intacta. A menos que sean deliberadamente incorporados en el diseño del estudio, los efectos de externalidades y cruce, debe reducirse al mínimo, o por lo menos, deberían ser documentados a fondo. (Ver como antecedente "**Las amenazas al diseño**")

1. Amenazas a la recolección de datos

a) Atrición

La atrición se produce cuando los evaluadores no reúnen información sobre las personas que fueron seleccionadas como parte de la muestra original. Nótese, que el grupo tratamiento y el grupo control, a través de la asignación aleatoria, se construyen para ser estadísticamente iguales al principio. El grupo de control tiene la intención de parecerse al contrafactual -lo que habría pasado al grupo de tratamiento si el tratamiento no hubiera sido ofrecido. (Ver: **¿Por qué Aleatorizar?**). Si las personas que abandonan el estudio son "idénticos" en ambos grupos de tratamiento y de control, es decir, si el grupo de control con menos personas aún representa un contrafactual válido para el grupo de tratamiento con menos personas, esto reduciría el tamaño de nuestra muestra, y podría truncar la población objetivo a la que nuestros resultados se pueden generalizar, pero no pondría en peligro la "verdad" de los resultados (al menos tal como se aplica a la población restringida).

Por ejemplo, supongamos que nuestra área de estudio es rural, y que muchos miembros del hogar pasan una parte significativa del año trabajando en zonas urbanas. Supongamos, además, que hemos creado nuestra muestra y recogido datos de línea base cuando

los miembros del hogar que migran estaban en casa durante las cosechas e incidentalmente para nuestro estudio. Si recogemos nuestros datos de medición final durante la temporada baja, los miembros de la familia que emigran habrán regresado a la ciudad y por tanto no estarán disponibles para nuestro estudio. Asumiendo que estos son los mismos individuos en los grupos de tratamiento y de control, nuestro estudio ahora se limita únicamente a los no migrantes. Si la población de no migrantes en el grupo control representa un buen contrafactual de la población no migrante en el grupo de tratamiento, nuestras estimaciones de impacto será perfectamente válidas, pero sólo aplicables a la población no migrante.

Sin embargo, si esa disminución tiene una forma distinta en los dos grupos, y los individuos restantes del grupo de control ya no sirven como un buen contrafactual, esto podría influir en nuestros resultados. Utilizando el ejemplo de las enfermedades transmitidas por el agua, supongamos que en el grupo de control más hijos y madres están enfermos. Como resultado, los jóvenes que suelen emigrar a las ciudades durante la temporada baja, se quedan en casa para ayudar a su familia. Los hogares que fueron asignados al grupo control contienen más inmigrantes en nuestra encuesta final. Los datos demográficos de los grupos de tratamiento y de control son ahora diferentes (mientras que en un principio, estaban equilibrados). Es factible que estos migrantes jóvenes sean típicamente más saludables. Ahora, a pesar de que nuestro tratamiento tuvo éxito en mejorar la salud de los niños y las madres, en promedio, nuestro grupo de control contiene a los trabajadores migrantes más saludables. Cuando se mide la incidencia de diarrea, los resultados de los inmigrantes sanos del grupo control podrían compensar por los resultados de sus familiares enfermos. Luego, al comparar los grupos de tratamiento y de control, no veríamos ningún efecto en absoluto y podríamos concluir que el tratamiento fue ineficaz. Este resultado sería falso y engañoso.

En este ejemplo simplificado, podríamos reintroducir el equilibrio mediante la eliminación de nuestra muestra de todos los migrantes. Con frecuencia, sin embargo, las características que podrían identificar de forma fiable a aquellos individuos que desaparecen, tanto reales como futuros, no han sido medidas, o son imposibles de observar. La predicción de atrición puede ser tan difícil como predecir la participación en los experimentos no aleatorios. Similarmente, el sesgo de atrición puede ser tan perjudicial como el sesgo de selección al hacer inferencias causales.

2. Externalidades y Efecto Control-Tratado

Las externalidades ocurren cuando individuos en el grupo de control son de alguna manera u otra afectados por el tratamiento. Por ejemplo, si ciertos niños están en el grupo de control de un estudio de entrega de pastillas de cloro, pero juegan con niños que están en el grupo de tratamiento, ahora tendrán amigos que tienen menos probabilidades de estar enfermos y por tanto tienen menos probabilidades de ellos mismos estar enfermos. En este caso, están indirectamente afectados por el programa, aunque hubieran sido asignados al grupo de control. Los individuos que "cruzan" son aquellos controles que encuentran la manera de ser directamente tratados. Por ejemplo, si la madre de un niño del grupo control lo lleva a beber agua del tanque de abastecimiento de un hogar en el grupo tratamiento, ella se está infiltrando en el grupo de tratamiento. Cumplimiento imparcial es un término más amplio, que comprende a aquellos que "cruzan" y también aquellos individuos tratados que deliberadamente deciden no participar (o poner cloro en el agua, en este ejemplo).

Cuando un estudio sufre de externalidades y efecto control-tratado, en muchos casos todavía es posible usar técnicas estadísticas para producir resultados válidos. Sin embargo, estos vienen con ciertos supuestos, muchos de los cuales estábamos tratando de evitar cuando optamos por una aleatorización en primer lugar. Por ejemplo, si las externalidades se pueden predecir con el uso de variables observadas, pueden ser controladas. Con el cumplimiento imparcial, si suponemos que los que no cumplieron no se vieron afectados por la intervención, y por la misma razón, las personas que "cruzaron" se vieron afectadas en la misma forma que los miembros del grupo de tratamiento, podemos inferir el impacto de nuestro programa. Pero como se explica en la sección **¿Por qué Aleatorizar?**, al hacer mas supuestos, el terreno en el que nos encontramos es menos firme a la hora de declarar que la intervención causo los resultados medidos.

CÓMO OBTENER RESULTADOS

Al final de una intervención (o por lo menos el periodo de evaluación de la intervención), los datos de la encuesta final deben ser recolectados para medir resultados finales. Asumiendo que la integridad de la asignación aleatoria se mantuvo y que la recolección

de datos se administró correctamente, es hora de analizar los datos. El método mas simple es medir el resultado promedio para el grupo de tratamiento y compararlo con el resultado promedio del grupo de control. La diferencia representa el impacto del programa. Para determinar si el impacto es estadísticamente significativo, uno puede testear la igualdad de promedios, usando un test-t simple. Uno de los muchos beneficios de las evaluaciones aleatorias es que el impacto puede ser medido sin la necesidad de técnicas estadísticas avanzadas. También se pueden realizar análisis más complicados. Por ejemplo, regresiones agregando controles para otras características para aumentar la precisión. Sin embargo, cuando se comienza a aumentar la complejidad del análisis, el número de potenciales errores también incrementa. Por tanto, el evaluador debe tener el conocimiento necesario y ser muy cauteloso al desempeñar este tipo de análisis.

Es importante notar que cuando se obtiene un resultado, no hemos "encontrado la verdad" con una certeza del 100 por ciento. Hemos producido un resultado que está cercano a la verdad, con cierto grado de probabilidad. Mientras más grande sea el tamaño de nuestra muestra (mas pequeños serán nuestros errores estándar y) tendremos más certeza. Sin embargo, nunca podemos tener una certeza del 100%.

Este hecho conlleva a dos tropiezos muy comunes durante el análisis:

1) **Resultados Múltiples**: La aleatorización no garantiza que el impacto estimado es perfectamente preciso. El impacto estimado no tiene sesgo, pero sigue siendo una estimación. El "azar" genera un margen de error alrededor de la verdad. Con bastante frecuencia, la estimación estará muy cerca de la verdad. Ocasionalmente, la estimación se desviará un poco más. En raras ocasiones, se apartará significativamente. Si usamos una medida de resultado, hay alguna posibilidad de que se haya desviado significativamente de la verdad. Pero esto es muy poco probable. Si estamos interesados en varios indicadores de resultados, muchos estarán cerca, pero otros se desviarán. Mientras más indicadores de resultado usemos, es más probable que uno o más se desvíen significativamente. Por ejemplo, supongamos que las pastillas de cloro que se distribuyen para combatir aquellas enfermedades transmitidas por el agua, estaban defectuosas o nunca se utilizaron. Si se comparan veinte diferentes indicadores de resultados, es muy probable que para alguno de ellos surgiera una mejora significativa en salud, y para otro una desmejora significativa. Si es que usamos suficientes indicadores de impacto, eventualmente vamos a encontrar uno que es significativamente distinto entre los grupos de tratamiento y control. Esto no es un problema en sí. El problema surge cuando el evaluador hace "data mining", buscando todos los indicadores de resultados, hasta que encuentra un impacto significativo, reporta este único resultado, y no presenta los otros resultados insignificantes que fueron descubiertos durante la búsqueda.

2) **Análisis de sub-grupos**: De modo similar, así como un evaluador puede hacer "data mining" al mirar muchos indicadores de resultados, el evaluador también puede hallar un resultado significativo al mirar a distintos subgrupos en aislamiento. Por ejemplo, puede ser que las pastillas de cloro no tengan ningún impacto aparente en la salud de los hogares. Podría ser razonable mirar el impacto en niños en el hogar, o en niñas en particular. Pero podríamos estar tentados a comparar niños y niñas de distintos grupos de edad, de distintas composiciones de hogares, en distintas combinaciones. Podríamos descubrir que hay mejoras significativas en la salud del grupo de tratamiento de niños entre 6 y 8 años, que sólo tienen una hermana, cuyo abuelo vive en el hogar y donde el hogar posee una televisión y ganado. Hasta podríamos inventar una historia plausible de por qué este subgrupo podría haber sido afectado y no otros. Pero si encontramos que éste es el único impacto significativo después de una serie de impactos insignificantes para otros subgrupos, es probable que la diferencia hubiera sido causada por el "azar" – no por nuestro programa.

COMO SACAR IMPLICANCIAS PARA POLÍTICAS PÚBLICAS

Tras realizar una evaluación aleatoria perfecta y un análisis de resultados honesto, podemos derivar implicaciones para políticas publicas con cierto nivel de certeza acerca de nuestras conclusiones de como el programa impacta nuestra población objetivo. Por ejemplo, "Nuestro programa de distribución de pastillas de cloro causó una reducción en la incidencia de diarrea en los niños en nuestra población objetivo en 20 puntos porcentuales". Esta declaración es científicamente legítima, o *válida a nivel interno*. El rigor de nuestro estudio no puede decirnos, sin embargo, si este mismo programa tendría el mismo impacto si se replicara en una población objetivo diferente, o si se expandiera. A diferencia de la validez interna, que puede ser proporcionada por una

evaluación aleatoria bien realizada, la validez externa, o generalización, es más difícil de obtener. Para extrapolar cómo nuestros resultados se aplicarían a un contexto diferente, tenemos que salir de nuestro rigor científico, y comenzar a responder en supuestos. Dependiendo de nuestro conocimiento del contexto de nuestra evaluación y de otros contextos en los que nos gustaría generalizar los resultados, nuestras suposiciones pueden ser más o menos razonables.

Sin embargo, la metodología que elegimos -una evaluación aleatoria- no provee validez interna a costa de la validez externa. La validez externa es una función del diseño del programa, los proveedores de servicios, los beneficiarios, y el entorno en que se llevó a cabo la evaluación del programa. Los resultados de cualquier evaluación de programas están sujetos a esas mismas realidades contextuales cuando se utilizan para obtener conclusiones respecto a programas similares o a las políticas aplicadas en otros lugares. Lo que la evaluación aleatoria nos compra es la certeza de que nuestros resultados al menos son válidos internamente.