



Foto: Paul Smith | J-PAL/IPA

MAYO DE 2024

# UNA INTRODUCCIÓN A LAS EVALUACIONES DE IMPACTO ALEATORIZADAS

**Preparado por: Shantal Aragón, Gaby Bustamante, Vianney Fernández, María Paz Monge y Valentina Olivares.**

**Edición general: María Paz Monge**

---

**Agradecimientos:** Agradecemos a Jeanne Lafortune, Claudia Macías, Sandra Peralta y Consuelo Sotomayor por sus valiosos comentarios. Todos los errores son nuestros.

*Este documento fue preparado por personal de J-PAL LAC en mayo de 2024. El propósito del documento es introducir brevemente las evaluaciones de impacto aleatorizadas y no representa una revisión exhaustiva del tema.*

## PREFACIO

Para garantizar el éxito de las políticas y programas implementados por los gobiernos, es crucial contar con información adecuada para la toma de decisiones. En este sentido, la evaluación juega un papel fundamental en la política pública, ya que permite identificar qué funciona y qué no, mejorando el diseño de futuras intervenciones y optimizando el uso de recursos públicos. Existen varios tipos de evaluación, cada uno con ventajas y desventajas. En este documento mencionaremos brevemente algunos, pero nos centraremos principalmente en uno que ha ganado relevancia durante los últimos años: **la evaluación de impacto aleatorizada**.

# CONTENIDOS

<b>1. INTRODUCCIÓN</b>	<b>4</b>
1.1 Sobre esta guía	4
1.2 El Ciclo de Aprendizaje para la cocreación de políticas y programas	5
<b>2. ENTENDIENDO CÓMO QUEREMOS GENERAR CAMBIOS Y MEDIR IMPACTO</b>	<b>8</b>
2.1 Teoría de Cambio	8
2.2 Indicadores para el monitoreo y evaluación del programa	11
2.3 Tipos de evaluación de programas	12
<b>3. EVALUANDO IMPACTO</b>	<b>15</b>
3.1 Métodos de evaluación de impacto	15
3.2 La importancia de seleccionar el método adecuado	17
<b>4. LAS EVALUACIONES DE IMPACTO ALEATORIZADAS</b>	<b>19</b>
4.1 La evaluación aleatorizada paso a paso	19
4.2 ¿Por qué aleatorizar?	22
4.3 ¿Cuándo (no) implementar una evaluación aleatorizada?	23
<b>5. CONSIDERACIONES ÉTICAS AL IMPLEMENTAR EVALUACIONES ALEATORIZADAS</b>	<b>25</b>
5.1 Consideraciones éticas en las ciencias sociales	25
5.2 Consideraciones éticas en las evaluaciones aleatorizadas	27
<b>6. DESAFÍOS PRÁCTICOS EN UNA EVALUACIÓN ALEATORIZADA</b>	<b>29</b>
6.1 Desafíos en el diseño	30
6.2 desafíos asociados a la implementación de un programa	34
<b>7. UTILIZANDO LOS RESULTADOS DE EVALUACIONES EN POLÍTICA PÚBLICA</b>	<b>37</b>
7.1 Caminos de la evidencia a la acción	37
7.2 Aplicando evidencia a nuestro contexto	38
7.3 Expandiendo un programa	43
<b>8. CASO DE ESTUDIO: EVALUACIÓN ALEATORIZADA DE PRINCIPIO A FIN</b>	<b>45</b>
<b>REFERENCIAS</b>	<b>48</b>
<b>GLOSARIO</b>	<b>51</b>

# 1. INTRODUCCIÓN

## 1.1 SOBRE ESTA GUÍA

Las instituciones de gobierno diseñan e implementan políticas y programas que llegan a millones de personas, con objetivos tan variados como aumentar la recaudación fiscal, mejorar las tasas de vacunación o disminuir la incidencia de la violencia dentro del hogar. Para que una nueva intervención consiga los resultados esperados, es clave que durante su formulación se tome en cuenta lo aprendido en experiencias similares, replicando lo que ha resultado exitoso y evitando lo que no.

La evaluación—entendida como el proceso por el cual se emiten juicios valorativos sobre las actividades y resultados de una política pública, un programa social, una estrategia o un proyecto—es clave al momento de formular políticas y programas. Al identificar qué funciona (y que no), la evaluación permite **mejorar el diseño de programas** y dirigir los recursos a aquellas estrategias con una mayor efectividad, **optimizando el gasto público**. Asimismo, el **conocimiento generado** en la evaluación de un programa puede utilizarse para el diseño de otro. Por ejemplo, si una evaluación muestra que un sistema de incentivos aumentó la participación del personal de salud, ese hallazgo se puede utilizar al momento de diseñar un esquema de incentivos para el personal de los establecimientos educacionales.

Existen diferentes tipos de evaluación en el contexto de programas y políticas públicas, cada uno con ventajas y desventajas que los hacen más adecuados para una u otra situación. Algunas evaluaciones se realizan para diseñar la intervención y otras para saber si esta fue efectiva; algunas se centran en verificar que el programa se esté implementado de acuerdo a lo planeado, mientras que otras se enfocan en los resultados del mismo; algunas se realizan en forma prospectiva y otras en forma retrospectiva; y así sucesivamente.

En esta guía nos enfocamos en presentar los principales conceptos relacionados con un tipo específico de evaluación: las **evaluaciones de impacto aleatorizadas**. El término “impacto”, se refiere a que la evaluación busca medir el cambio atribuible al programa, mientras que el término “aleatorizada” se incluye para indicar la forma específica en que se mide el impacto. Además de las evaluaciones aleatorizadas, existen otros métodos para evaluar impacto que, si se implementan correctamente, pueden generar evidencia sumamente rigurosa.

## RECUADRO 1: MONITOREO VS. EVALUACIÓN

Por una parte, el **monitoreo** es un proceso continuo de recopilación y análisis de información sobre un programa. Durante el monitoreo, se comparan los resultados reales con los previstos para juzgar **qué tan bien se está implementando la intervención**. Por otra parte, la **evaluación** mide diferentes elementos del programa para determinar el **valor o relevancia del programa**, con el objetivo de proporcionar información creíble para que quienes toman decisiones identifiquen formas de lograr los resultados deseados.



**Recurso de profundización.** El documento “[Basic Principles of Monitoring and Evaluation](#)” de la Organización Internacional de Trabajo ofrece la definición previa y ahonda en los principios de monitoreo y evaluación.

## 1.2 EL CICLO DE APRENDIZAJE PARA LA COCREACIÓN DE POLÍTICAS Y PROGRAMAS

La incorporación de evidencia y evaluaciones en la política pública no es trivial, pero puede ser operacionalizada a través del llamado “Ciclo de Aprendizaje” (ver Figura 1). El Ciclo de Aprendizaje es una herramienta conceptual útil para la cocreación de políticas públicas y programas sociales. Principalmente, ayuda a las organizaciones a identificar cuándo y cómo usar la evidencia en las diferentes etapas del proceso de diseño e implementación de políticas. Una de sus ventajas radica en que es flexible a las condiciones y capacidades de cada gobierno u organización. El Ciclo de Aprendizaje tiene tres fases que se describen a continuación.

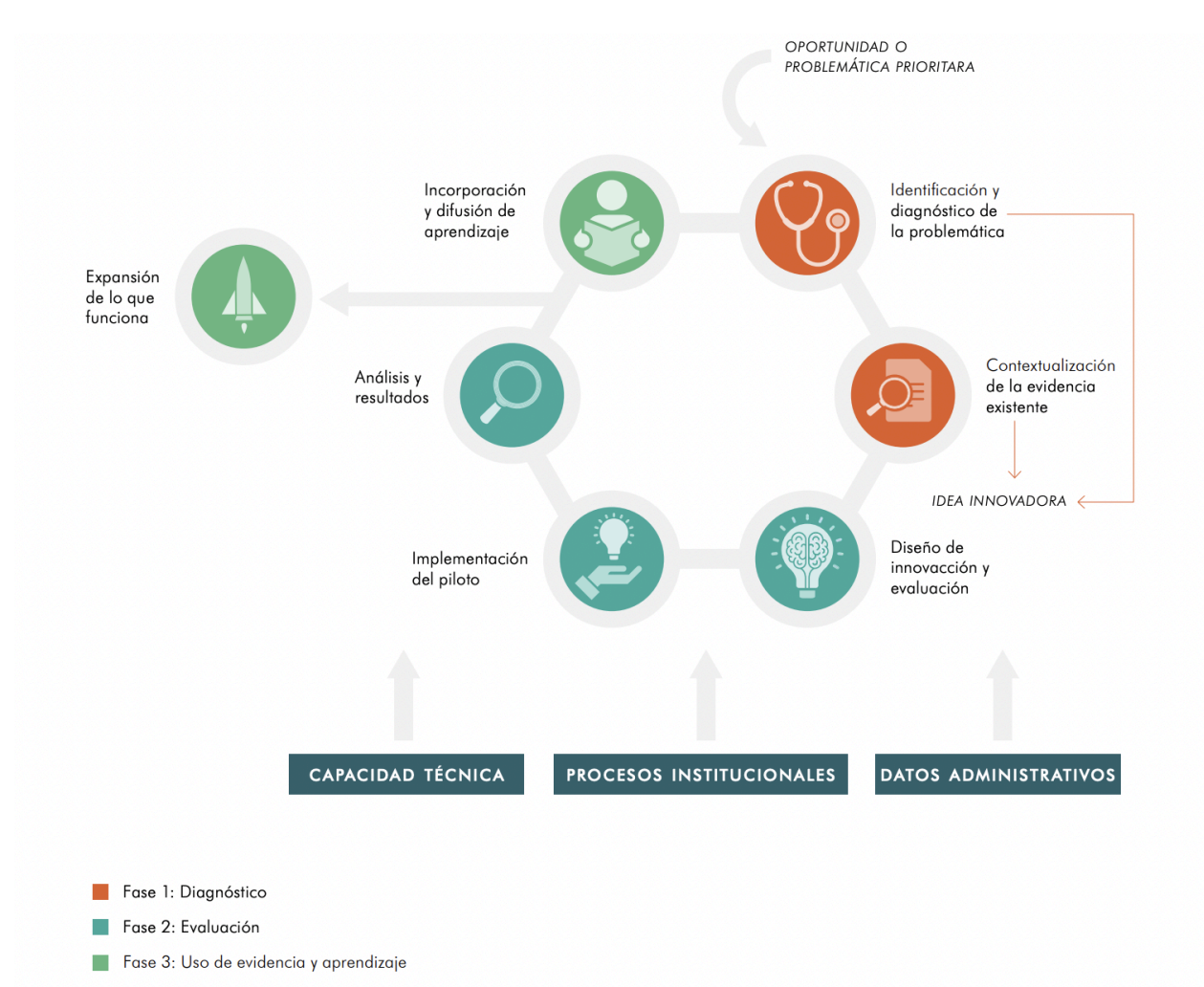
**Fase 1: Diagnóstico.** En primer lugar, la organización diagnostica la naturaleza, alcance y posibles causas del problema utilizando datos, aprovechando su conocimiento del contexto local y apoyándose en personas expertas y equipos de investigación. Luego, revisa la evidencia para explorar potenciales soluciones. Si existe evidencia rigurosa que indica que un programa determinado podría funcionar, el programa generalmente se adapta al contexto local y se pilotea a pequeña escala. De lo contrario, se puede diseñar y pilotear un nuevo programa a partir de lo que se ha aprendido de otras experiencias. Cuando el programa piloteado sigue un modelo que ya ha sido evaluado con resultados positivos, la agencia de gobierno puede expandirlo o escalarlo, cuidando que los mecanismos principales del programa original se mantengan; si no se dispone de evidencia suficiente sobre el programa propuesto, es clave evaluar su impacto antes de tomar una decisión sobre la continuidad y escala de la intervención.

**Fase 2: Evaluación.** Cuando el programa piloteado sigue un modelo que ya ha sido evaluado con resultados positivos, la agencia de gobierno puede expandirlo o escalarlo sin necesidad de hacer una evaluación de impacto. Sin embargo, la implementación y escalamiento podrían requerir una evaluación de procesos para verificar que la implementación del programa a escala está llevándose a cabo de la forma esperada. Si no se dispone de evidencia suficiente sobre el impacto del programa propuesto, es clave evaluar el mismo antes de tomar una decisión sobre la continuidad y escala de la intervención. La Fase 2 del Ciclo de Aprendizaje consiste en diseñar e implementar una evaluación de impacto para estimar si el programa tuvo los efectos esperados y comprender los mecanismos que lo hicieron posible (o no). Es recomendable que la evaluación la haga un equipo de investigación externo, pero que la organización gubernamental se involucre fuertemente en el proceso.

**Fase 3: Uso de evidencia y aprendizaje.** Los gobiernos toman en consideración diferentes factores para decidir sobre sus programas, entre ellos los resultados de la evaluación. Si la evaluación concluye que el programa tuvo un impacto positivo y el impacto justifica los costos del programa, el gobierno tendrá argumentos a favor de expandir el programa. Por otro lado, si la evaluación indica que el programa no solucionó el problema identificado en el diagnóstico, existirán argumentos a favor del rediseño, ajuste o reducción del programa. En caso de ser necesario, el programa rediseñado puede someterse a otra ronda de diagnóstico y evaluación, que comprenda las fases 1 y 2. Es importante que se compartan públicamente los resultados de la evaluación para que otras organizaciones que enfrentan problemas y contextos de políticas similares puedan informarse de las lecciones aprendidas.

Aunque el concepto del Ciclo de Aprendizaje pueda parecer simple, en la realidad, cada gobierno se enfrenta a condiciones iniciales, capacidades y limitaciones únicas. Además, la formulación de políticas rara vez sigue un proceso tan claro y lineal como el esquematizado en el Ciclo de Aprendizaje. No todos los gobiernos tienen la capacidad o la necesidad de implementar completamente el Ciclo de Aprendizaje en su funcionamiento, sino que pueden aplicarlo poco a poco, por ejemplo, comenzando por ciertas áreas o proyectos.

Figura 1. El Ciclo de Aprendizaje



 **Recurso de profundización.** El Ciclo de Aprendizaje se adapta a las condiciones, capacidades y particularidades de cada institución. Consulte ejemplos su aplicación en el [Informe Forjando una Cultura para el Uso de Evidencia: Lecciones de J-PAL sobre sus Alianzas con Gobiernos en Latinoamérica.](#)

## 2. ENTENDIENDO CÓMO QUEREMOS GENERAR CAMBIOS Y MEDIR IMPACTO

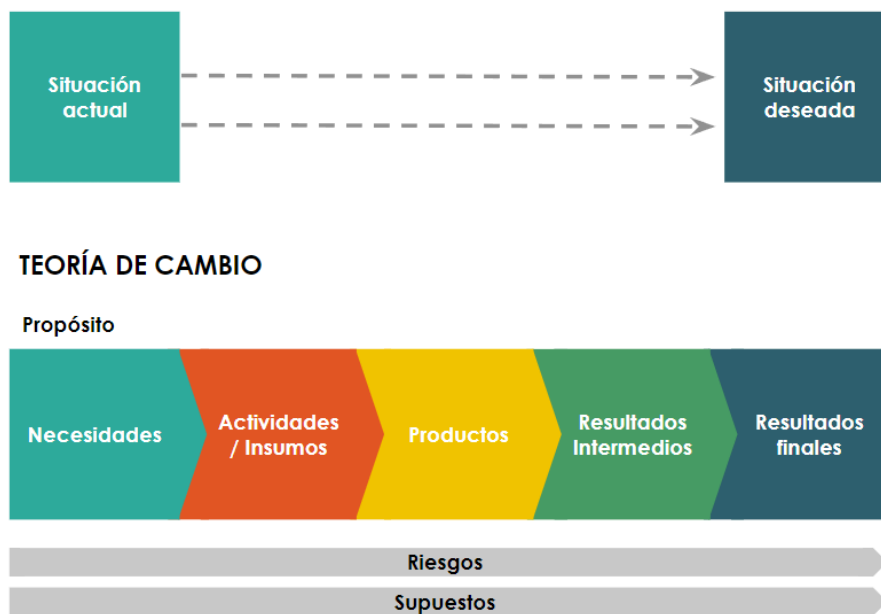
Antes de evaluar el impacto de un programa, tenemos que entender muy bien cómo planeamos generar ese impacto. Debemos conocer la situación actual, definir lo que queremos lograr y determinar cómo llegaremos allí. Y solo una vez que tengamos claro lo anterior podremos evaluar si efectivamente logramos el cambio esperado.

### 2.1 TEORÍA DE CAMBIO

La Teoría de Cambio es una **metodología que ayuda a comprender cómo generar un cambio**. Concretamente, consiste en representar gráficamente la ruta que debemos recorrer: dónde nos encontramos hoy (necesidades identificadas), dónde queremos llegar (resultados deseados) y cómo pensamos llegar al destino..

La Teoría de Cambio indica cuál es la **lógica causal de la intervención**, es decir, la relación causa y efecto entre las acciones y resultados del proyecto. Así, podemos identificar con claridad qué mecanismos de un programa son los que conducen a los resultados finales (ver Figura 2). Por esto, **la Teoría de Cambio cumple un rol fundamental tanto en el monitoreo como en la evaluación de un programa**.

**Figura 2.** La Teoría de Cambio explicita cómo llegar a la situación deseada



La Teoría de Cambio se compone de varios elementos. En primer lugar, el **propósito** corresponde al objetivo macro del programa. En segundo lugar, las **necesidades, actividades, productos, resultados intermedios y resultados finales** indican el recorrido que haremos desde la situación actual hasta la deseada. Estos cinco elementos se ordenan según una lógica causal: cada uno de ellos influye en el elemento siguiente. Finalmente, los **riesgos** y los **supuestos** son consideraciones transversales que juegan un papel importante en el éxito del programa.

- **Propósito.** Objetivo general y macro que queremos lograr con la intervención. Puede pensarse como “la razón por la que existe el programa” y se asemeja a la estructura de la misión o visión de una organización.
- **Necesidades.** Punto de partida a la cadena lógica o causal. Se refiere a las necesidades específicas que se quieren abordar en un contexto dado y debe centrarse en la población objetivo.
- **Actividades e insumos.** Recursos clave que se requieren para construir los productos del programa y las acciones principales que deben ejecutarse para implementar el proyecto.
- **Productos.** Bienes y servicios entregados por el programa directamente a los participantes. Resultan de la combinación entre actividades e insumos, y aunque a veces pueden confundirse con una reformulación de las actividades, el énfasis está en el servicio recibido.
- **Resultados Intermedios.** Cambios de corto plazo que la intervención produce en los participantes y que tienen que ver con cambios en sus *actitudes, conocimientos, capacidades y/o comportamientos*. Normalmente, es uno de los enfoques principales de las evaluaciones de impacto. Los resultados intermedios son instrumentales, es decir, sirven como herramientas para lograr los resultados finales, que son el objetivo en sí mismo.
- **Resultados Finales.** Último eslabón de la cadena causal. Son los cambios a largo plazo en el estado de la población objetivo y son consecuencia de los resultados intermedios. Los resultados intermedios consideran un plazo menor que los resultados finales.
- **Supuestos.** Condiciones externas que deben cumplirse para que la cadena causal sea válida. Son situaciones que no dependen de los implementadores, pero que son necesarias para que el programa funcione con éxito. Es importante pensar en medidas de mitigación en caso de que no tengamos certeza de que algún supuesto se va a cumplir.
- **Riesgos.** No son impedimentos a la implementación del programa, sino que son consecuencias negativas generadas por el programa, de forma colateral. Un programa puede estar alcanzando los efectos positivos esperados, y al mismo tiempo estar produciendo efectos no deseados en otros ámbitos. Debemos generar medidas de mitigación para los riesgos y evaluar si es que efectivamente el programa tuvo consecuencias negativas.



**Recurso de profundización.** Esta [presentación](#) de J-PAL explica en mayor detalle la Teoría de Cambio (en inglés).

## RECUADRO 2: UNA TEORÍA DE CAMBIO EN UN PROGRAMA DE TUTORÍAS ESCOLARES

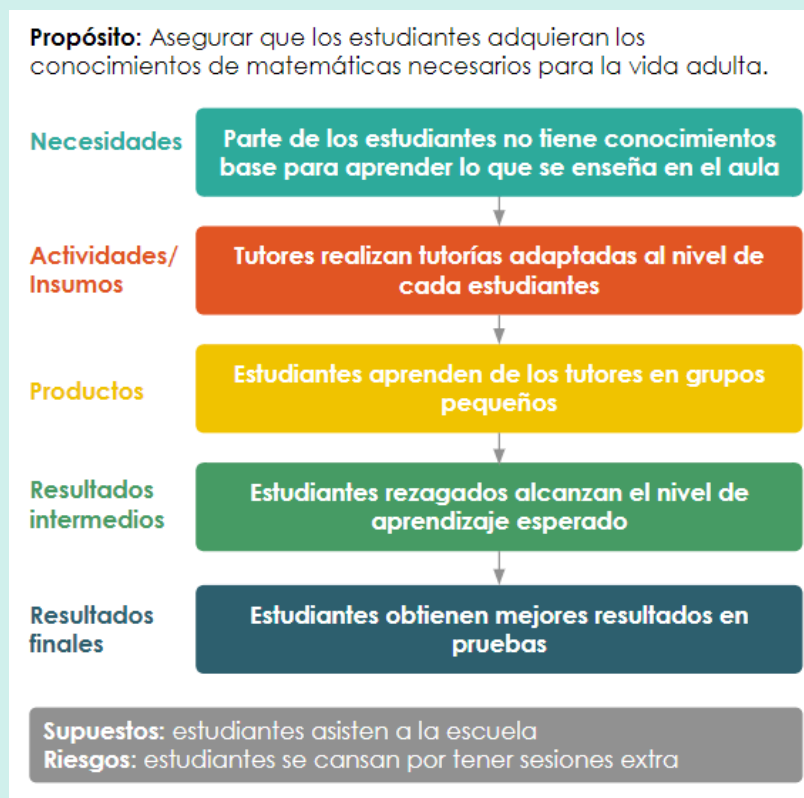
La Teoría de Cambio se entiende mejor con un ejemplo. Imaginemos que tenemos un programa de tutorías para estudiantes de 2° grado. En ese caso, nuestra Teoría de Cambio se vería como la de la ilustración abajo (este ejemplo es una versión simplificada con fines ilustrativos).

En primer lugar, el **propósito** es asegurar que los estudiantes adquieran los conocimientos y habilidades en matemáticas necesarios para la vida adulta.

En segundo lugar, tenemos la cadena causal. La **necesidad** es que existe una proporción importante de estudiantes que no han alcanzado el nivel de aprendizaje esperado para su curso, lo que les dificulta aprender nuevos contenidos. Por ejemplo, si en 2° grado se enseña la resta, un niño que no sabe sumar va a tener

dificultades para aprenderla. Dada esa necesidad, las **actividades e insumos** consisten en clases extra en grupos pequeños para aquellos estudiantes que aún no dominan los contenidos fundamentales. Al trabajar en grupos pequeños adaptados al nivel de aprendizaje de cada niña y niño (**producto**), los estudiantes aprenderán los contenidos que no dominaban. Así, el **resultado intermedio** es que los estudiantes rezagados alcanzarán el nivel esperado para su grado y podrán aprender nuevos contenidos en las clases regulares. Y esto se verá reflejado en el largo plazo en que los estudiantes tendrán mejores resultados en las pruebas (**resultado final**).

Finalmente, es necesario considerar los supuestos y riesgos. Ya que el programa se enfoca en lo que acontece al interior del establecimiento educacional, estamos asumiendo que los estudiantes van a la escuela (**supuesto**). La asistencia de los estudiantes es una condición necesaria para el éxito del programa, pero escapa al control de este. Por otra parte, existe el **riesgo** de que los estudiantes se agobien con las tutorías, por lo que debemos buscar formas de que las tutorías no sean una carga extra.



### RECUADRO 3: PERSPECTIVA DE GÉNERO EN EL DISEÑO Y EVALUACIÓN DE PROGRAMAS

Los programas y políticas pueden tener impactos diferentes en mujeres y hombres. Por ejemplo, un estudio en India observó que otorgar un préstamo con un período de gracia aumentó las ganancias empresariales de los hombres, pero en promedio no tuvo ningún efecto sobre las ganancias empresariales de las mujeres ([Bernhardt et al. 2019](#)). El equipo de investigación decidió indagar más profundamente y notó que en los hogares con un solo negocio, los negocios de mujeres y hombres obtuvieron rendimientos similares. En cambio, en los hogares con múltiples negocios, las mujeres obtuvieron menores rendimientos, lo que se explicaría en parte por que ellas redirigían parte del préstamo a los negocios de sus maridos.

Este caso ilustra la **importancia de utilizar una perspectiva de género para entender el impacto que tendrá una intervención sobre las mujeres y tomar medidas al respecto**. Por otra parte, también es importante asegurar que la evaluación del programa incorpore **datos de género** que permitan visualizar tanto las barreras que enfrentan las mujeres como los efectos que tiene el programa sobre dicha población.



**Recurso de profundización.** La [Guía práctica para la medición del empoderamiento de las mujeres y las niñas en evaluaciones de impacto](#) de J-PAL ofrece consejos prácticos para medir el empoderamiento.

## 2.2 INDICADORES PARA EL MONITOREO Y EVALUACIÓN DEL PROGRAMA

Una vez graficada la Teoría de Cambio, tenemos una referencia clara para el monitoreo del programa, tanto en términos de la implementación, como de los resultados o impacto esperados. Así, es posible ir confirmando la ocurrencia de cada eslabón de la cadena causal según lo planificado en la Teoría de Cambio.

La Teoría de Cambio y sus elementos pueden ser monitoreados a través de uno o más indicadores. Los indicadores asociados a los insumos, actividades y productos, conforman los **indicadores de gestión**. Los indicadores de gestión se evalúan a través de una evaluación de procesos, y permiten determinar qué tan adecuada ha sido la implementación del programa, según lo planificado.

Por otro lado, los **indicadores de resultados** intermedios y finales permiten hacer dos tipos de análisis diferentes:

- **Evaluación de resultados.** Corresponde al proceso de medición y análisis de las variables sobre las cuales el programa busca influir o generar un cambio, es decir, sobre las variables de

resultados (intermedios o finales). Este tipo de evaluaciones puede implicar, por ejemplo, hacer mediciones antes y después del programa, para ver cómo varió en el tiempo.

- **Evaluación de impacto.** Permite determinar el efecto *generado por el programa* sobre las mismas variables objetivo (resultados intermedios o finales). A diferencia de la evaluación de resultados, cuantifica la diferencia que es generada exclusivamente por el programa, aislándola de otros factores que influyen sobre el resultado. Como se menciona más adelante, la evaluación del impacto de un programa, requerirá aplicar metodologías estadísticas de inferencia causal.

Aunque para realizar una evaluación de impacto no es estrictamente necesario contar con indicadores de gestión, es altamente recomendable hacerlo, ya que estos ayudan a comprender las razones detrás de la efectividad del programa. Por ejemplo, si un programa no tiene impacto, pero según nuestros indicadores no se llevaron a cabo todas las actividades planeadas, no significa que el programa diseñado no sirva, sino que la implementación no se realizó en forma adecuada. En cambio, si no tenemos indicadores de gestión, no sabemos si el problema está en el diseño del programa o si fallamos en ejecución.

Al momento de seleccionar qué indicadores vamos a monitorear, debemos tomar en consideración cómo vamos a obtener dicho indicador (por ejemplo, a través de data administrativa, encuestas, data de gestión, etc.) y con qué frecuencia.

#### RECUADRO 4: INDICADORES EN UN PROGRAMA DE TUTORÍAS ESCOLARES

Volviendo al ejemplo del Recuadro 2, debemos tener indicadores para cada uno de los elementos de la Teoría de Cambio, por ejemplo:

- **Necesidades:** porcentaje de estudiantes que saben sumar y restar.
- **Actividades/insumos:** cantidad de sesiones de tutorías implementadas.
- **Productos:** porcentaje de asistencia a tutorías por estudiante.
- **Resultados intermedios:** notas durante el semestre.
- **Resultados finales:** puntaje en matemáticas en una prueba estandarizada en 3° grado y en grados superiores.
- **Supuestos:** porcentaje de asistencia a la escuela por estudiante.
- **Riesgos:** porcentaje de estudiantes que declara estar cansado en encuesta de fin de año.

## 2.3 TIPOS DE EVALUACIÓN DE PROGRAMAS

Aunque esta guía se enfoca en la evaluación aleatorizada, es importante mencionar que existen varios tipos de evaluación de programas y conocer las diferentes opciones permite decidir cuál es mejor en cada contexto (ver Figura 3). A continuación se revisan brevemente algunos tipos de evaluación.

- **Evaluación de necesidades.** Es un análisis de la situación en el que identificamos la población con la que trabajamos y diagnosticamos la naturaleza, el alcance y las causas del problema o necesidad que queremos abordar. Se realiza antes de la implementación de un programa.
- **Evaluación teórica.** Análisis de los planteamientos que sustentan el programa y cómo se espera que la intervención impacte a la población objetivo, con el fin de evaluar la viabilidad y factibilidad del proyecto desde una perspectiva teórica. Existen diversas herramientas para realizarla, siendo muy populares el Marco Lógico y la Teoría de Cambio. Se realiza antes de la implementación de un programa.
- **Evaluación de procesos.** Evaluación que determina si el programa se implementa de acuerdo a cómo se planificó. Utiliza una variedad de métodos, tanto cuantitativos como cualitativos, para medir el progreso durante la implementación.
- **Evaluación de resultados.** Seguimiento de las variables sobre las cuales el programa busca influir o generar un cambio, pero no distingue lo que puede ser directamente atribuible al programa de lo que no.
- **Evaluación de impacto.** Evaluación que determina la magnitud, cuantitativa y cualitativa, del cambio en la población que podemos atribuir directamente al programa (y no a otros factores). Es un proceso puntual, realizado una vez considerando un período limitado en el tiempo. Se diseña previamente a la intervención y los resultados se obtienen posteriormente.
- **Evaluación de costo-efectividad.** Análisis para estimar cuánto cuesta generar el impacto deseado en la población. Se puede realizar después de la implementación del programa comparando el impacto del programa con su costo. También, se puede realizar en forma prospectiva, utilizando los resultados de una evaluación de impacto similar y calculando los gastos esperados del programa.

Los diferentes tipos de evaluaciones se complementan entre sí. Por ejemplo, la evaluación de necesidades sirve para construir una evaluación teórica y la evaluación teórica a su vez ayuda a determinar los indicadores de las evaluaciones de procesos y resultados. Asimismo, no podemos llevar a cabo una evaluación de costo efectividad si no tenemos una evaluación de impacto que estime la efectividad del programa.

**Figura 3.** Tipos de evaluaciones y ejemplos

RESUMEN	TIPO DE PREGUNTAS QUE RESPONDE
<p>La <b>evaluación de necesidades</b> se centra en el contexto y tiene busca determinar si el programa aborda un problema real y relevante, y cuáles son las causas de dicho problema.</p>	<ul style="list-style-type: none"> <li>¿Cuál es la naturaleza y la magnitud del problema?</li> <li>¿Cuáles son las posibles causas?</li> <li>¿Cuáles son las características de la población objetivo?</li> <li>¿Qué programas son los más pertinentes?</li> <li>¿Qué programas ya existentes abordan el problema?</li> </ul>
<p>La <b>evaluación teórica</b> se enfoca en la teoría detrás de cómo se pretende resolver el problema identificado, ofreciendo una narrativa detallada del programa.</p>	<ul style="list-style-type: none"> <li>¿Es viable la solución planteada?</li> <li>¿Cuáles son los resultados finales que se esperan lograr?</li> <li>¿Cómo se pretende llegar a esos resultados finales?</li> <li>¿Se implementó un programa similar en otros lugares?</li> <li>¿Responde el programa a las necesidades?</li> </ul>
<p>La <b>evaluación de procesos</b> evalúa si la implementación del programa se llevó a cabo de acuerdo con la planificación establecida, asegurando la conformidad con el relato propuesto.</p>	<ul style="list-style-type: none"> <li>¿Se implementa el programa según lo planeado?</li> <li>¿Cuántas personas están recibiendo el servicio? ¿Son las personas correctas?</li> <li>¿Es adecuado el servicio en términos de cantidad y calidad?</li> <li>¿Cuenta el personal del programa con todas las competencias requeridas?</li> <li>¿Cómo se administran los recursos?</li> </ul>
<p>La <b>evaluación de resultados</b> da seguimiento de las variables de interés, sin distinguir lo que es directamente atribuible al programa de lo que no.</p>	<ul style="list-style-type: none"> <li>¿Cómo han cambiado los indicadores de interés?</li> <li>¿Se alcanzaron los resultados esperados?</li> <li>¿Qué cambios experimentó la población objetivo?</li> </ul>
<p>La <b>evaluación de impacto</b> proporciona información sobre si se alcanzaron los objetivos del programa y en qué medida se lograron.</p>	<ul style="list-style-type: none"> <li>¿Qué efectos tuvo el programa? ¿Se mantuvieron en el tiempo?</li> <li>¿Son necesarios todos los componentes del programa?</li> <li>¿Algunos grupos de personas fueron más afectadas por la intervención que otros?</li> <li>¿Existen efectos adversos no planificados?</li> </ul>
<p>La <b>evaluación de costo-efectividad</b> determina si los impactos se obtuvieron de manera eficiente y con un uso adecuado de los recursos disponibles.</p>	<ul style="list-style-type: none"> <li>¿Es el costo del programa razonable en relación con la magnitud del impacto?</li> <li>¿Existen intervenciones alternativas que cumplirían los mismos objetivos a menor costo?</li> <li>¿Los recursos son usados de manera eficiente?</li> </ul>

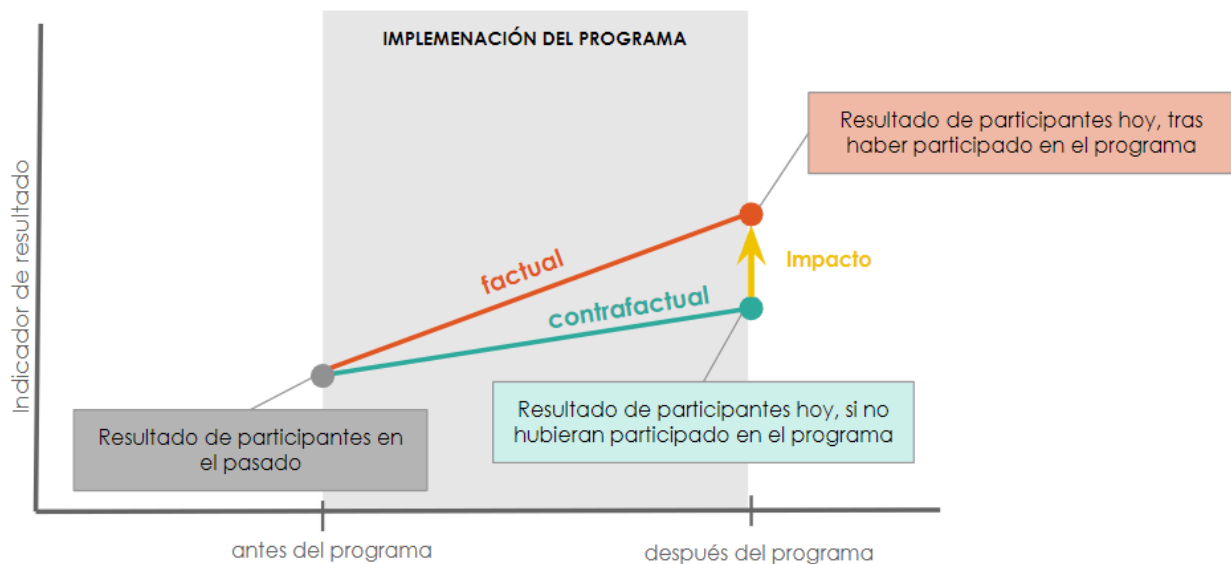
### 3. EVALUANDO IMPACTO

#### 3.1 MÉTODOS DE EVALUACIÓN DE IMPACTO

Tal como se mencionó en la Sección 2, las evaluaciones de impacto buscan medir el **efecto causal** del programa, es decir, cuáles fueron los **cambios en los indicadores de resultados que podemos atribuir directamente al programa (y no a otros factores)**. Básicamente, las evaluaciones de impacto buscan comparar lo que pasó al implementarse el programa (factual) con *lo que hubiera pasado si el programa nunca se hubiera implementado* (contrafactual).

La Figura 4 ilustra esto. El eje horizontal representa el paso del tiempo y el eje vertical los resultados. La línea naranja corresponde al factual, es decir, cómo cambiaron los resultados después de implementar el programa. La línea verde corresponde al contrafactual, es decir, cómo hubieran cambiado los resultados *si no se hubieran implementado el programa*. Para conocer los cambios atribuibles al programa (impacto), es necesario comparar resultado obtenido tras el programa (punto naranja) con el resultado que si hubiera obtenido sin el programa (punto verde).

**Figura 4.** Representación gráfica de la evaluación de impacto de un programa



Dado que es imposible que simultáneamente un programa se implemente y no se implemente, **se debe encontrar una forma de simular el contrafactual**, es decir, el escenario hipotético en que el programa no existe. La **principal diferencia entre los diversos métodos de evaluación de impacto radica en cómo se simula el contrafactual con el que se compara el programa**. Por ejemplo, si se comparan los participantes antes y después del programa, el contrafactual serían los resultados de los participantes

antes del programa y el factual serían los resultados de esas mismas personas después del programa. La Figura 5 presenta diferentes métodos de evaluación de impacto y cuál es el contrafactual que utilizan. La validez técnica de cada método dependerá de la rigurosidad del diseño de la investigación.

**Figura 5.** Algunos métodos de evaluación de impacto

MÉTODO	DESCRIPCIÓN
Pre-post o antes-después (método no experimental)	Compara los resultados de los participantes del programa antes y después del programa.
Diferencias simples (método no experimental)	Compara los resultados de los participantes del programa con los resultados de personas que no participaron en el programa.
Diferencias en diferencias (método no experimental)	Compara la evolución (antes y después) de los resultados de quienes participaron en el programa, con la evolución de los que no participaron. Es decir, compara los cambios que hubo entre quienes participaron y no (en términos relativos)
Regresión discontinua (método cuasi-experimental)	Este método se puede utilizar cuando la elegibilidad para participar en el programa está determinada por un puntaje de corte (por ejemplo, solo pueden participar quienes tengan un ingreso de menos de \$100 o un puntaje de vulnerabilidad mayor a 80 puntos). En este caso se comparan los resultados de las personas que están justo por debajo del puntaje de corte (y que quedaron fuera del programa) con las que están justo por encima.
VARIABLES INSTRUMENTALES (método cuasi-experimental)	El diseño utiliza una "variable instrumental" que predice la participación en el programa. Luego, el método compara los resultados de individuos según su participación prevista, en lugar de su participación real <sup>1</sup> .
Evaluación aleatorizada (método experimental)	Determina aleatoriamente quienes participan y no en el programa y compara los resultados de ambos.

<sup>1</sup> Por ejemplo, si se quiere estimar el impacto de la propaganda electoral digital sobre la participación en las elecciones presidenciales es difícil saber si alguien consume o no propaganda digital. En este caso se podría utilizar el acceso a internet como variable instrumental y comparar los resultados de las personas con banda ancha con los resultados de las personas sin banda ancha.





*Revisar cada uno de los métodos de evaluación de impacto escapa del objetivo de esta guía, no obstante es importante destacar que los métodos de evaluación cuasi-experimentales bien diseñados y ejecutados pueden entregar resultados robustos y en algunos casos ser más adecuados que una evaluación aleatorizada.*



**Recurso de profundización.** Este [documento](#) de J-PAL resume varios tipos de evaluación de impacto, indicando los supuestos en que se basan y la data que requieren (en inglés).

## 4. LAS EVALUACIONES DE IMPACTO ALEATORIZADAS

En los últimos años, las evaluaciones aleatorizadas<sup>2</sup>, han adquirido cada vez más importancia como herramienta para medir el impacto de políticas públicas. En 2019, los cofundadores de J-PAL Abhijit Banerjee y Esther Duflo, y el profesor afiliado a J-PAL Michael Kremer, recibieron el [Premio Nobel de Economía](#) por promover este método de investigación que ha transformado la política social y el desarrollo económico.

Las evaluaciones aleatorizadas permiten obtener una estimación rigurosa del impacto causal de una intervención. Es decir, estimar qué cambios específicos en la vida de las personas participantes se pueden atribuir al programa. Este método de evaluación también permite **responder preguntas específicas sobre la efectividad de un programa y su teoría de cambio**: ¿qué tan efectivo fue este programa?, ¿hubo efectos secundarios no deseados?, ¿quiénes se beneficiaron más?, ¿qué componentes del programa funcionaron o no?, ¿qué tan costo-efectivo fue el programa?, ¿podemos agregar componentes adicionales para aumentar la efectividad?

Adicionalmente, una evaluación aleatorizada, puede **responder otras preguntas que permiten mejorar el programa y diseñar nuevas intervenciones**: ¿qué componentes del programa se podrían replicar en otro contexto?, ¿cómo se compara este programa con otros diseñados para lograr objetivos similares?, ¿es más conveniente invertir en este programa o en otra alternativa?



**Recurso de profundización.** Este [documento](#) de J-PAL ahonda en las ventajas de las evaluaciones aleatorizadas (en inglés).

### 4.1 LA EVALUACIÓN ALEATORIZADA PASO A PASO

Las evaluaciones aleatorizadas son un tipo de evaluación de impacto, es decir, buscan **medir los cambios en la población objetivo que pueden ser atribuidos al programa evaluado** (ver la Sección 3). Al igual que todos los métodos de evaluación de impacto, las evaluaciones aleatorizadas buscan comparar lo que realmente pasó con el programa (factual) con *lo que hubiera pasado si el programa no existiera* (contrafactual). Dado que es imposible que un programa exista y simultáneamente no exista, el contrafactual es una situación hipotética que nunca podemos observar en la realidad.

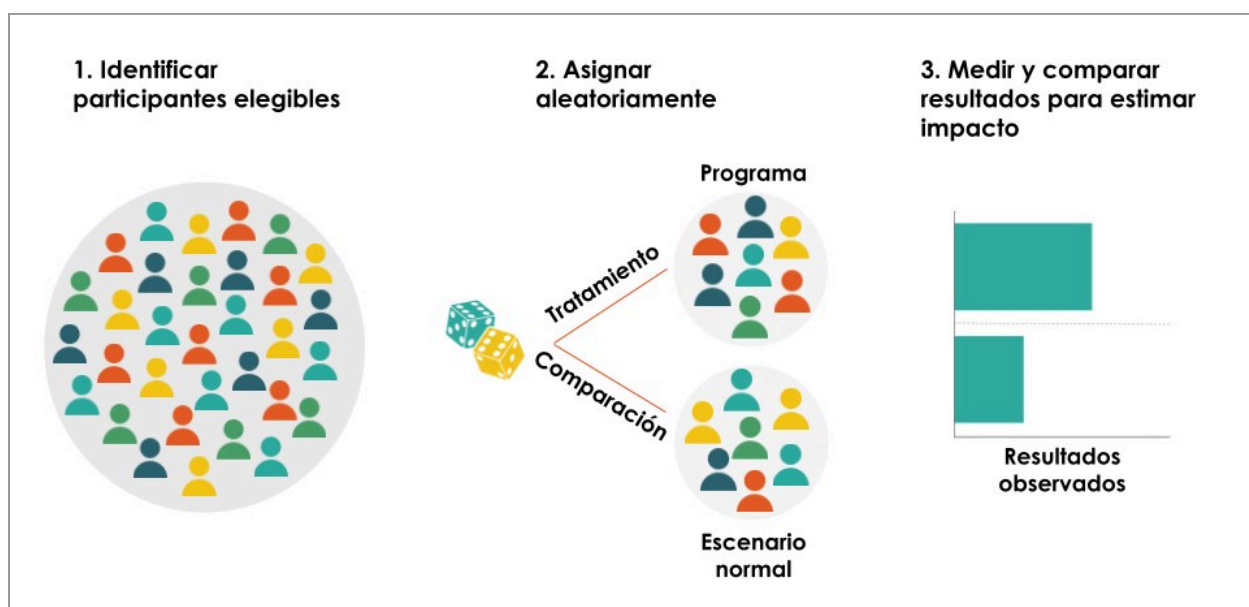
Como el contrafactual no se puede observar, cada metodología de evaluación de impacto construye o simula el contrafactual de una forma diferente. En las evaluaciones aleatorizadas, esto se hace a través de la **asignación aleatoria**. En términos prácticos, esto quiere decir que quienes son parte del estudio se asignan al azar a uno de dos grupos:

<sup>2</sup> “Randomized controlled trials” (o “RCT” por sus siglas en inglés)

- **Grupo de tratamiento.** Grupo que participa en el programa<sup>3</sup>. Este grupo representa al factual dentro de la evaluación aleatorizada.
- **Grupo de comparación<sup>4</sup>.** Grupo que no participa en el programa durante la evaluación, pero que puede participar en el programa más tarde. Este grupo representa al contrafactual que simulamos a través de la evaluación aleatorizada.

A diferencia de algunos otros métodos de evaluación de impacto, **una evaluación aleatorizada se debe diseñar antes de que el programa sea implementado.** La Figura 7 esquematiza una evaluación aleatorizada y sus diferentes pasos.

**Figura 7.** La evaluación aleatorizada paso a paso

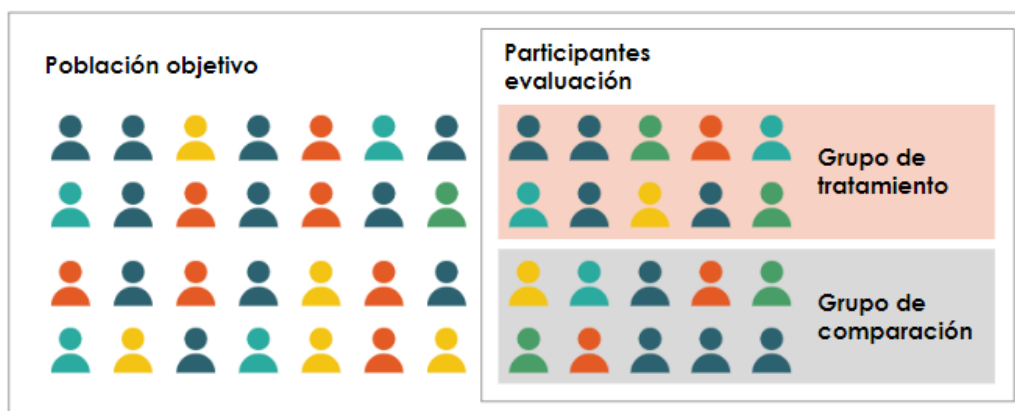


<sup>3</sup> Una evaluación puede buscar evaluar más de un programa o diferentes variantes de un programa, en ese caso se tendrían varios grupos de tratamiento.

<sup>4</sup> También denominado "grupo de control" o "grupo de referencia".

En primer lugar, debemos identificar quiénes participarán en la evaluación. Es importante notar que los participantes de una evaluación no corresponden necesariamente a la población completa que es elegible para recibir el programa, sino que en muchos casos son un subconjunto de esta (ver Figura 8). Por ejemplo, aunque un programa de tutorías esté diseñado para que lo reciban todos los estudiantes, el programa podría evaluarse solo con los estudiantes de una ciudad y si los resultados son positivos, ampliar el programa a todo el país.

**Figura 8.** En la evaluación participa un subconjunto de la población objetivo



Una vez que determinamos quiénes participarán en la evaluación, asignamos aleatoriamente algunos participantes al grupo de tratamiento y a otros al grupo de comparación. La aleatorización se puede hacer en diferentes niveles. Por ejemplo, podemos seleccionar al nivel de personas (como trabajadores o estudiantes), a nivel de organizaciones (como firmas o escuelas), a nivel de unidad geográfica (como municipios o barrios), entre otros.

El Recuadro 5 ahonda en qué tomar en cuenta al momento de seleccionar el nivel en el que aleatorizaremos, pero **como mínimo debemos aleatorizar al nivel de la unidad en la que se entregará el tratamiento**. Por ejemplo, si nuestro programa consiste en ofrecer un bono para las empresas que contratan personas en situación de discapacidad, no tiene sentido aleatorizar al nivel de personas, ya que este es un incentivo que recibe la firma y todas las personas dentro de esa firma serán afectadas por el programa.

## RECUADRO 5: LA UNIDAD DE ALEATORIZACIÓN

La unidad de aleatorización se refiere a qué es lo que asignaremos aleatoriamente al grupo de tratamiento o de comparación. Por ejemplo, para evaluar un programa de capacitaciones laborales podríamos aleatorizar al nivel de trabajadores (¿quiénes reciben la capacitación?) o al nivel de firma (¿qué empresas tienen acceso a las capacitaciones?)

Normalmente, la unidad de aleatorización es la unidad de observación, es decir, la unidad para la que recolectamos información. Volviendo a nuestro ejemplo anterior, si recopilamos información sobre los ingresos de los individuos, probablemente vamos a aleatorizar a nivel de trabajador, mientras que si nos enfocamos en la productividad de la empresa, lo más natural es que aleatoricemos las firmas.

En algunos casos, la unidad de aleatorización contiene múltiples unidades de observación. Esto se llama aleatorización por conglomerados. Con la aleatorización por conglomerados, las unidades de observación se agrupan en conjuntos (conglomerados), y la aleatorización ocurre a nivel del conglomerado en lugar de a nivel de la unidad.

Varias consideraciones relacionadas con la validez del experimento influyen en la elección de una unidad de aleatorización. Asimismo, la unidad de aleatorización afectará el tamaño de la muestra necesario para la evaluación. En general, al subir de nivel en la aleatorización—por ejemplo, al pasar de aleatorizar a nivel de trabajadores a aleatorizar al nivel de firmas— se requieren más unidades de observación para la evaluación. **Como mínimo, la unidad de aleatorización debe ser la unidad en la que se entregará el tratamiento.**

El último paso consiste en estimar el impacto. Una vez que el programa es implementado, debemos medir los indicadores de interés—como los puntajes en pruebas estandarizadas, ingreso del hogar, ventas, etc.—tanto en el grupo de tratamiento como en el grupo de comparación<sup>5</sup>. La diferencia entre los resultados de ambos grupos es el impacto del programa.

## 4.2 ¿POR QUÉ ALEATORIZAR?

Al comparar dos grupos cabe preguntarnos si esos grupos son realmente comparables<sup>6</sup>. Por ejemplo, si ofrecemos un programa de microcrédito a personas que ya tenían un negocio y comparamos sus ingresos con los de un grupo de personas que no tienen un negocio, no sabremos qué parte de las diferencias se deberán al microcrédito y qué parte al hecho de tener un negocio previo.

<sup>5</sup> Aunque no siempre es necesario, en muchos casos es conveniente medir también los indicadores antes de la implementación del programa.

<sup>6</sup> En términos coloquiales, no queremos comparar peras con manzanas.

La Figura 7 ayuda a comprender mejor esto. Imaginemos que los participantes de nuestro estudio son como los del círculo más de la izquierda y que los colores representan diferentes características. Por ejemplo, las personas naranjas podrían ser quienes ya tienen un negocio, las verdes ser aquellas que han participado en capacitaciones empresariales, las azules las que recién comenzarán a emprender y así sucesivamente. Para hacer una comparación justa, queremos que los integrantes del grupo de tratamiento sean similares a los del grupo de comparación, es decir que en ambos grupos haya personas que ya tienen un negocio, algunas que hayan participado en capacitaciones, unas cuantas que recién comiencen a emprender, etc.

La asignación aleatoria es el elemento clave de una evaluación aleatorizada, ya que nos permite construir dos grupos comparables. Conceptualmente, **la aleatorización significa que cada unidad tiene la misma probabilidad de ser asignada a un grupo determinado**. Por ejemplo, si la persona A tiene 40 por ciento de probabilidades de ser asignada al grupo de tratamiento y 60 por ciento de probabilidades de ser asignada al grupo de comparación, la persona B también tiene que tener 40 y 60 por ciento de probabilidades de ser asignada al tratamiento y comparación, respectivamente. Si tenemos una cantidad suficientemente grande de participantes en la evaluación y los asignamos aleatoriamente a los grupos de tratamiento y comparación, antes de la intervención ambos grupos serán similares. Es decir, en los dos grupos tendremos una proporción semejante de participantes con cada una de las características<sup>7</sup>. Esto se observa gráficamente en los dos círculos en medio de la Figura 2, donde vemos que la proporción de personas de cada color es similar en el grupo de tratamiento y de comparación<sup>8</sup>.

**Al utilizar la aleatorización, la única diferencia entre los grupos de tratamiento y control es si participan o no en el programa—y no sus características—, por lo que una vez que la intervención se implemente podemos atribuir las diferencias entre grupos al impacto del programa.**

### 4.3 ¿CUÁNDO (NO) IMPLEMENTAR UNA EVALUACIÓN ALEATORIZADA?

Las evaluaciones aleatorizadas son excelentes herramientas para aprender sobre el impacto de una política o programa, pero no son apropiadas para todas las situaciones. Al momento de decidir si llevar a cabo o no una evaluación aleatorizada debemos tomar en consideración diferentes aspectos:

1. **Cantidad de participantes.** Si bien la cantidad de participantes que necesitamos para tener una buena evaluación va a variar de un caso a otro (ver Recuadro 6), se necesita que un número considerable de personas participen en la evaluación. Por ello, los programas con pocos usuarios son malos candidatos para ser evaluados mediante una evaluación aleatorizada.

---

<sup>7</sup> Esto es cierto para características medibles u observables (como la edad, la nacionalidad y el nivel socioeconómico) y para características más difíciles de medir no observables (como la motivación, la resiliencia y la tolerancia a la frustración).

<sup>8</sup> La composición de los grupos no tiene que ser necesariamente idéntica: decimos que dos grupos son estadísticamente equivalentes cuando en promedio sus características no difieren en forma sistemática.

2. **Posibilidad de aleatorizar.** Existen casos en que no se puede hacer una asignación aleatoria, como con una política macroeconómica o cuando es éticamente inaceptable restringir el acceso a un grupo de comparación.
3. **Costos.** Los costos de implementar una evaluación aleatorizada deben compararse con los beneficios de la información que esta nos entregará. No obstante, la creciente disponibilidad de datos administrativos puede disminuir considerablemente los costos de una evaluación.
4. **Plazos.** Una evaluación aleatorizada toma tiempo y debe diseñarse antes de la implementación o ampliación del programa evaluado. Si el programa ya se implementó o si se necesitan resultados de forma urgente, puede ser mejor optar por otros métodos.
5. **Viabilidad política.** Una evaluación aleatorizada exitosa requiere el involucramiento de quienes implementan la política pública y es necesario tomar en cuenta la resistencia que una evaluación podría generar.
6. **Integridad del diseño.** Se debe analizar si es posible asegurar la integridad del diseño de investigación (por ejemplo, si podemos evitar que las personas del grupo de tratamiento y de comparación no se cambien de grupo).
7. **Evidencia existente.** En los casos en que existe suficiente evidencia sobre la efectividad de un mecanismo o programa, es mejor enfocarse en tener una buena implementación que en realizar una nueva evaluación.

#### RECUADRO 6: PODER Y CANTIDAD DE PARTICIPANTES

Incluso si es que un programa es efectivo, puede darse el caso que una evaluación aleatorizada no capture el efecto del programa. En este contexto, el **poder** de una evaluación se refiere a la probabilidad de que la evaluación detecte el impacto de un programa, siempre y cuando el programa realmente tenga un efecto. Existen varios factores que determinan el poder de una evaluación, entre ellos la cantidad de personas o unidades que participan en la evaluación: como regla general, mientras más unidades participen en la evaluación, esta tendrá un mayor poder.



**Recurso de profundización.** J-PAL ofrece [recursos](#) para calcular el poder de una evaluación (en inglés)

## 5. CONSIDERACIONES ÉTICAS AL IMPLEMENTAR EVALUACIONES ALEATORIZADAS

Los equipos de investigación suelen enfrentar situaciones complejas que implican dilemas éticos o morales; dependiendo del contexto, la percepción de lo que es moralmente aceptable o justo suele variar de formas inesperadas.

Por ello, es importante que quienes están participando en la conducción de una evaluación aleatorizada tengan una actitud proactiva, anticipando soluciones a potenciales dilemas y se muestren reactivos para responder apropiadamente ante situaciones imprevistas. A continuación, presentamos algunas prácticas para obtener el máximo beneficio de la investigación y el mínimo riesgo para todas las personas participantes<sup>9</sup>.

### 5.1 CONSIDERACIONES ÉTICAS EN LAS CIENCIAS SOCIALES

En 1978, como consecuencia de las atrocidades cometidas contra la población afroamericana en el Estudio Tuskegee sobre sífilis y otras investigaciones científicas deshonestas hechas por científicos Nazis, el gobierno de los Estados Unidos formó una comisión especial para identificar las directrices que deben regir la investigación médica y social que involucre la participación de seres humanos. A las principales conclusiones de la comisión se les conoce como el **Informe Belmont**<sup>10</sup>.

El informe señala tres pilares fundamentales para la práctica ética de la investigación científica (ver Figura 9):

- **Respeto por las personas.** Las personas deben ser tratadas como agentes autónomos y que tienen algo que decir respecto a su participación en el estudio y la información será utilizada en el estudio. Además, reconoce que algunas personas pueden no ejercer plenamente su autonomía y requieren de medidas adicionales para tomar una decisión informada sobre su participación en una investigación; este es el caso de las personas que no han alcanzado la mayoría de edad<sup>11</sup>, privadas de la libertad y/o en situación de vulnerabilidad.
- **Beneficencia.** El diseño de un estudio que involucra la participación de humanos debe maximizar el beneficio de la investigación y minimizar riesgo para las personas participantes. La

---

<sup>9</sup> Esta sección busca introducir algunos conceptos importantes relacionados con la ética de la investigación, pero no es exhaustiva.

<sup>10</sup> Belmont Report en inglés.

<sup>11</sup> Cuando se trabaja con poblaciones que no pueden dar legalmente su consentimiento, además del consentimiento de la persona tutora, es importante contar la aprobación directa del sujeto de investigación. Por ejemplo, antes de obtener las medidas antropométricas de una persona que no ha alcanzado la mayoría de edad, se le debe preguntar si está de acuerdo con el contacto físico, en búsqueda de su consentimiento verbal.

participación de una persona en el estudio no debe implicar un riesgo superior al que implica la realización de cualquier otra actividad<sup>12</sup>.

- **Justicia.** Las personas o poblaciones que asumen el riesgo de participar en la investigación también deben recibir los beneficios de esta.

Además de quienes participan directamente en el estudio, es importante **tener en cuenta a todas las otras personas que de una u otra forma pueden verse implicados o afectados por este**. Por ejemplo, es importante asegurar condiciones adecuadas para que el equipo de investigación desempeñe su trabajo y que sus miembros estén preparados para abordar situaciones complicadas. Otro ejemplo es que debemos considerar los posibles efectos no deseados del estudio en las personas que no participaron directamente en la investigación<sup>13</sup>.

En la práctica, el equipo de investigación debe tomar medidas para que en todo momento la evaluación sea conducida bajo los máximos estándares éticos posibles. Además, **el diseño de las investigaciones debe ser revisado por un comité de ética**<sup>14</sup> que la evaluación se ciñen a los principios arriba mencionados.

**Figura 9.** En la evaluación participa un subconjunto de la población objetivo

RESPECTO POR LAS PERSONAS	BENEFICENCIA	JUSTICIA
Se debe reconocer la autonomía de las personas y proteger a aquellas con autonomía disminuida.	Se tiene la obligación de no hacer daño, acrecentar al máximo los beneficios y disminuir los daños posibles.	Las personas o poblaciones que participan en la investigación también deben recibir los beneficios de esta.



**Recursos de profundización.** El [Informe Belmont](#) describe en mayor profundidad los principios éticos para la protección de sujetos humanos en investigación y entrega directrices y consejos para su aplicación; la [Regulación Federal 45 CFR 46](#) del gobierno de los Estados Unidos fija directrices sobre cómo trabajar con poblaciones en situación de vulnerabilidad.

<sup>12</sup> Un concepto relacionado con la beneficencia es “equipoponderación” (equipoise) que, en el ámbito de las ciencias sociales, alude a la continuación del ejercicio de investigación hasta que sea claro cómo y en qué medida debe proveerse una intervención entre la población objeto de estudio.

<sup>13</sup> Por ejemplo, si una capacitación laboral dificulta el acceso al empleo a quienes no participaron de ella, el estudio debe medir tales efectos para determinar el impacto real de la intervención. En caso de que esto no sea posible, el equipo de investigación deberá reconsiderar la pertinencia del estudio.

<sup>14</sup> “IRB” por sus siglas en inglés (Institutional Review Board).

## 5.2 CONSIDERACIONES ÉTICAS EN LAS EVALUACIONES ALEATORIZADAS

Al hablar de evaluaciones aleatorizadas, es común preguntarse si es ético asignar personas al grupo de comparación, pues tal asignación podría implicar negarles el acceso a un bien o servicio. Al respecto, es importante tener en cuenta lo siguiente:

- **Una evaluación aleatorizada es ética cuando no se cuenta con pruebas rigurosas sobre la eficacia de una intervención.** En estas circunstancias, la evaluación es útil, pues genera evidencia para tomar una decisión informada sobre ampliar intervenciones efectivas o reasignar recursos de intervenciones ineficaces.
- **Es posible llevar a cabo una evaluación aleatorizada cuando la demanda para una intervención es muy alta y no se cuenta con recursos para cubrir a toda la población.** En este sentido, la aleatorización un método más justo de asignación de recursos que otros mecanismos como el orden de llegada<sup>15</sup>. En este caso, una evaluación aleatorizada puede cambiar el proceso de selección, pero no el número de personas que recibirán el bien o servicio provisto por la intervención.
- **Se puede realizar una evaluación aleatorizada sin negar el acceso al programa.** En intervenciones que buscan garantizar el acceso a un derecho (por ejemplo, a la justicia o a la educación gratuita), las evaluaciones aleatorizadas pueden orientar sobre la forma más eficiente de distribuir recursos a la población. En esa línea, una evaluación aleatorizada puede comparar la efectividad de dos versiones diferentes de una determinada intervención: la versión existente y una versión con una innovación o componente nuevo. Otra aproximación es no negarle el acceso a nadie, pero promover activamente el programa solo entre las personas del grupo de tratamiento; así la proporción de personas que participa en el programa será mayor en el grupo de tratamiento que en el grupo de comparación<sup>16</sup>.



**Recurso de profundización.** El sitio web de J-PAL tiene una [sección](#) que revisa en mayor detalle las consideraciones éticas de las evaluaciones aleatorizadas (en inglés).

<sup>15</sup> Un gran problema con ofrecer un programa por orden de llegada es que las personas que más lo necesitan no son necesariamente las que primero se inscriben.

<sup>16</sup> Este tipo de evaluación tiene lo que se conoce como “diseño de estímulo” (“encouragement design” en inglés); ver sección 6.1.3.

## RECUADRO 7: EVALUACIÓN ALEATORIZADA CUANDO NO SE CUENTA CON RECURSOS PARA ATENDER A TODA LA POBLACIÓN: TECHO

Un equipo de investigación realizó una [evaluación aleatorizada](#) para medir el impacto de la mejora de la infraestructura de vivienda en barrios marginales urbanos en El Salvador, México y Uruguay ofrecida por la ONG TECHO. Debido a las limitaciones presupuestarias y de personal, TECHO llevó a cabo loterías dentro de los asentamientos para seleccionar qué hogares recibirían soluciones habitacionales de emergencia (“mediaguas”). De 2.373 hogares elegibles en los tres países, 1.356 fueron seleccionados para recibir mejoras de vivienda y los 1.017 restantes sirvieron como grupo de comparación. Se concluyó que proporcionar mejores viviendas en los barrios marginales urbanos era relativamente económico y aumentaba sustancialmente la satisfacción con la vida en múltiples contextos, por lo que esta acción era una alternativa a la reubicación de las personas residentes en nuevas casas más alejadas de los centros urbanos y de los mercados laborales.

## 6. DESAFÍOS PRÁCTICOS EN UNA EVALUACIÓN ALEATORIZADA

Al llevar a cabo una evaluación aleatorizada se pueden presentar diversos desafíos. En esta sección revisamos algunos de los más comunes (ver resumen en la Figura 10) y planteamos posibles soluciones, diferenciando los problemas relacionados con el diseño de la evaluación de los que se originan durante la implementación del programa.

**Figura 10.** Algunos desafíos que pueden presentarse al realizar una evaluación aleatorizada

DESAFÍO	SOLUCIONES
<b>Existen recursos para proveer el programa a toda la población objetivo.</b> Si todos reciben el programa simultáneamente, no hay un grupo comparación.	Usar un diseño de fases: primero se entrega el programa a un grupo y después se va ampliando a otros grupos hasta cubrir a toda la población objetivo.
<b>El programa tiene criterios de elegibilidad estrictos;</b> si quienes los cumplen ya reciben el programa, puede no ser adecuado asignarlos al grupo de comparación.	Relajar un poco los criterios de elegibilidad y aleatorizar entre los que antes no eran elegibles.
<b>El programa es un derecho.</b> Por ello, no se puede crear un grupo de comparación sin acceso al programa.	Usar un diseño en que todos tienen acceso al programa, pero promover el programa solo entre los miembros del grupo de tratamiento.
<b>La cantidad de participantes es pequeña,</b> lo que disminuye el poder estadístico.	(1) Disminuir el nivel al que se hace la aleatorización (ej.: aleatorizar a nivel de hogar en vez de barrio) (2) Estratificar <sup>17</sup> de acuerdo a variables altamente correlacionadas con los resultados.
<b>Es difícil mantener la adherencia a la asignación aleatoria, debido a razones políticas o logísticas.</b> Si algunas unidades asignadas al grupo de comparación	(1) Asignar el grupo que recibe el programa y el de comparación a diferentes proveedores del servicio.

<sup>17</sup> Estratificar se refiere a subdividir la población en subgrupos de acuerdo a sus características. Por ejemplo, dividirlos según su nivel de ingresos, experiencia previa, conocimientos, etc.

DESAFÍO	SOLUCIONES
participan en el programa, puede haber “contaminación” entre grupos.	(2) Aumentar el nivel en que se hace la aleatorización (ej.: aleatorizar a nivel de barrio en vez de a nivel de hogar).
<b>El grupo de comparación se entera del programa, o es beneficiado/dañado por este.</b> Esto podría llevar a que haya “contaminación” entre grupos o que algunas unidades se retiren del estudio.	(1) Aumentar el nivel en que se hace la aleatorización (ej.: aleatorizar a nivel de barrio en vez de a nivel de hogar). (2) Crear “zonas de amortiguación”: evitar el contacto directo entre los miembros del grupo de tratamiento y de comparación.

## 6.1 DESAFÍOS EN EL DISEÑO

Ciertas características de un programa pueden implicar desafíos para utilizar una evaluación aleatorizada. A continuación revisaremos cuatro características que pueden dificultar el uso de una evaluación aleatorizada y cómo abordarlos.

### 6.1.1 El programa cuenta con recursos para cubrir a toda la población

Cuando un programa no cuenta con suficientes recursos para atender a toda la población objetivo, una asignación aleatoria puede ser un criterio justo para seleccionar quienes participarán en la intervención. Sin embargo, ¿qué hacer cuando el programa sí cuenta con recursos para atender a toda la población? ¿Cómo construir el grupo de comparación?

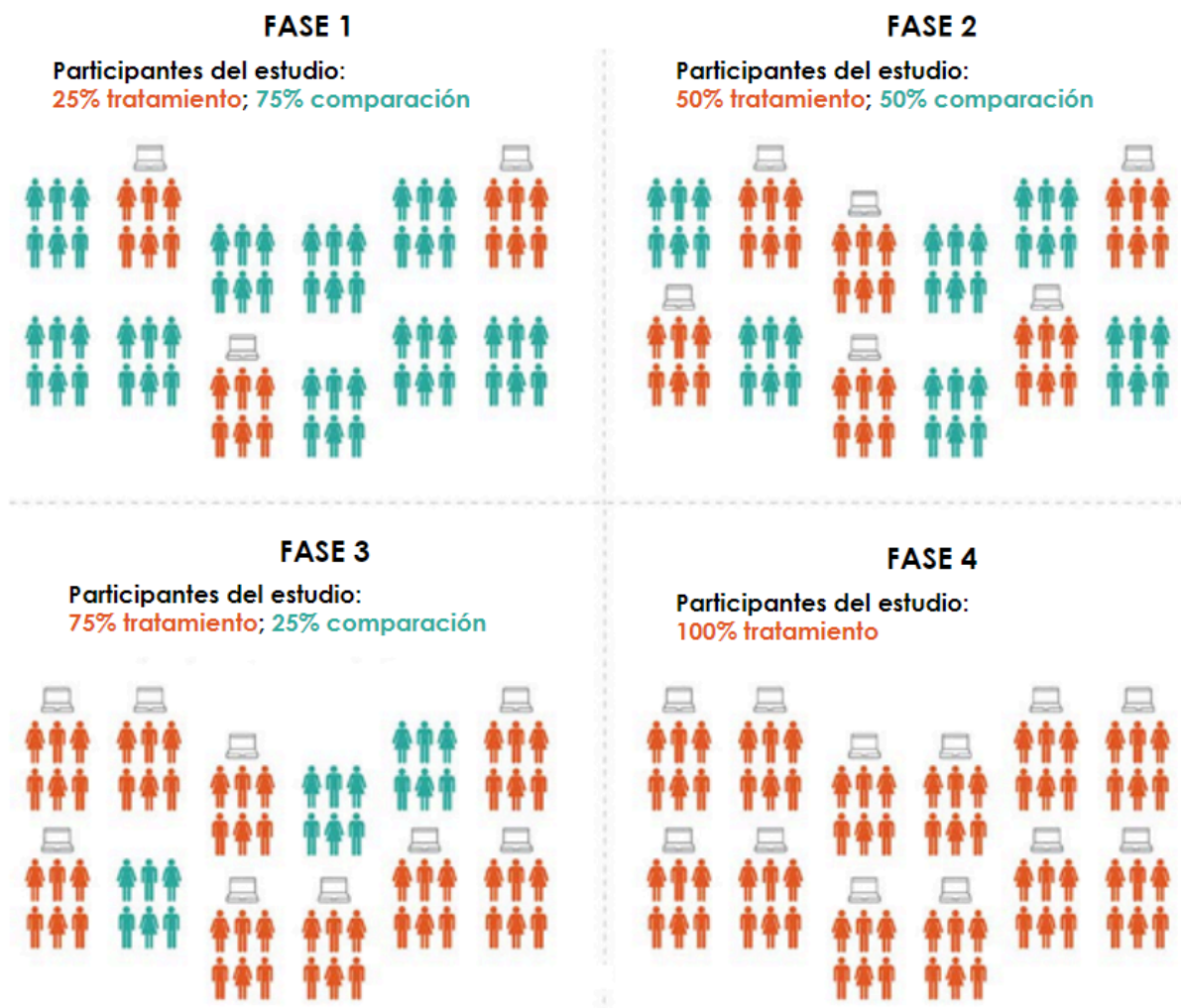
En los casos en que no tenemos sobredemanda podemos realizar **una aleatorización por etapas**<sup>18</sup>. En este diseño el grupo de tratamiento recibe primero el programa y el grupo de comparación los recibe solo después de que ha pasado un periodo de tiempo y se han medido los resultados<sup>19</sup>. Esto permite que todas las personas participantes se vean beneficiadas por la intervención sin perder el contrafactual en el corto plazo, pero limita la posibilidad de medir los efectos de largo plazo del programa<sup>20</sup>.

<sup>18</sup> “Phase-in design” en inglés.

<sup>19</sup> Este diseño puede incluir puede considerar más de dos etapas.

<sup>20</sup> En el largo plazo todos reciben el programa por lo que no tenemos un grupo de comparación

Figura 11. Aleatorización por etapas



### 6.1.2 El programa tiene criterios de selección estrictos

Muchos programas tienen criterios de selección estrictos que separan a las personas elegibles para recibir una intervención de quienes no lo son (por ejemplo, un ingreso máximo o la existencia de determinada precondition). En estos casos, la aleatorización entre la población que cumple con los criterios no es lo más adecuado, pues ya cuentan con el acceso al programa.

Una posible solución es **relajar un poco el criterio de selección y realizar la evaluación solo con quienes antes no eran elegibles y ahora sí**<sup>21</sup>. Así, este diseño responde la pregunta: “¿es efectivo el programa al expandirlo a la nueva población?”, pero tiene la limitante de que no determina si la

<sup>21</sup>Es importante notar que quienes ya cumplen con los criterios originales participan en el programa, pero no en la evaluación.

intervención es efectiva entre quiénes ya cumplían con los requisitos originales<sup>22</sup>. Otro problema de este diseño es que requiere de recursos adicionales para que el programa cubra a más personas.

#### RECUADRO 8: RELAJANDO LOS CRITERIOS DE SELECCIÓN DE UN PROGRAMA PARA EVALUARLO

Imaginemos que queremos evaluar el impacto de una transferencia monetaria que actualmente se ofrece a los hogares con un ingreso menor a \$100 mensuales. Una alternativa para evaluar este programa es mantener el beneficio para los hogares que tienen un ingreso menor a \$100, pero realizar la evaluación con quienes tienen un ingreso un poco mayor. Por ejemplo, podríamos identificar los hogares con ingresos entre \$100 y \$110 y asignar aleatoriamente algunos de ellos al grupo de tratamiento y a otros al grupo de comparación. Esto nos permitiría conocer el efecto del programa en los hogares con ingresos entre \$100 y \$110, pero no podremos generalizar directamente los resultados a los hogares con ingresos menores a \$100.

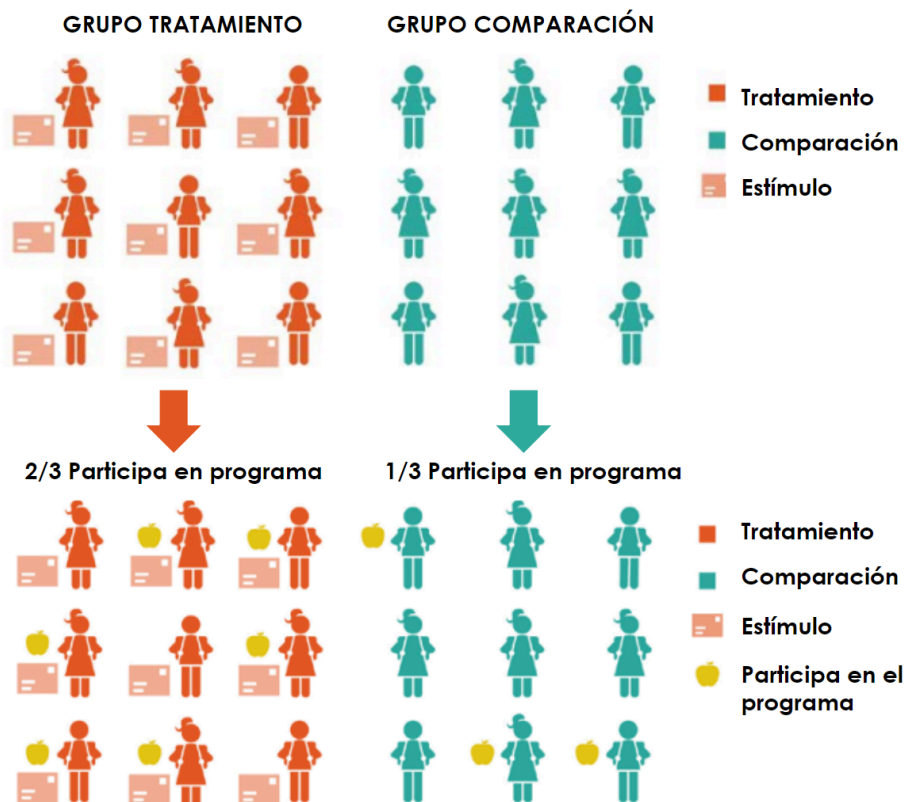
#### 6.1.3 El programa busca garantizar un derecho o un beneficio a un segmento poblacional determinado

Cuando se evalúa un programa que (1) **gran parte de la población no está aprovechando** y (2) en que **no se puede forzar o excluir la participación**, conviene utilizar un **diseño basado en estímulos**. En este diseño, tanto el grupo de tratamiento como el de comparación tienen acceso al programa, pero solo los miembros del primero reciben un estímulo para hacer uso del programa. El estímulo puede ser por ejemplo un incentivo monetario pequeño o una carta que recuerde a las personas que son elegibles para el programa y cuáles son los pasos a seguir para participar en él.

---

<sup>22</sup> Es decir, no es posible generalizar los resultados a las personas que ya estaban recibiendo el programa porque cumplían los requisitos originales.

Figura 12. Diseño basado en estímulos



Los estímulos deben diseñarse de manera tal que lleven a mayores tasas de participación en el grupo de tratamiento que en el de comparación. **Es importante señalar que una evaluación con este diseño, más que medir el impacto directo del programa, mide el impacto de recibir el estímulo.**

#### 6.1.4 La cantidad de participantes del estudio es pequeña

Una evaluación con una cantidad de participantes pequeña reduce la probabilidad de identificar un impacto, incluso cuando el programa sea efectivo<sup>23</sup>. Evaluaciones con insuficientes participantes no permiten concluir si el programa es o no efectivo, pudiendo implicar un desperdicio de recursos financieros y tiempo, proveer poca información útil o incluso, dañar la percepción de programas potencialmente efectivos.

Una forma de abordar el desafío de tener pocos participantes es **aleatorizar a un nivel inferior al que inicialmente se consideró**. Por ejemplo, es posible aleatorizar a nivel salón de clases, en lugar de aleatorizar por escuela o, aleatorizar por barrio, en lugar de comuna. Sin embargo, esta solución podría

<sup>23</sup> Técnicamente, esto implica que la evaluación tiene un poder estadístico bajo.

aumentar la probabilidad de que personas que no fueron asignadas al grupo de tratamiento se beneficien del programa, ensuciando los resultados de la evaluación<sup>24</sup>.

Otra medida que podemos tomar cuando la cantidad de participantes es pequeña es realizar una **asignación estratificada**. Esta técnica consiste en: (1) dividir a los participantes en subgrupos según características observables<sup>25</sup>; (2) dentro de cada subgrupo, asignar aleatoriamente a algunos participantes al grupo de tratamiento y a otros al grupo de comparación. La estratificación aumenta el poder estadístico de la evaluación y permite identificar el impacto del programa en subgrupos. Existen consideraciones técnicas al momento de estratificar: no debemos crear demasiados subgrupos (porque podrían quedar desbalanceados) y las características según las que estratificamos deben estar relacionadas con el resultado esperado<sup>26</sup>.

## 6.2 DESAFÍOS ASOCIADOS A LA IMPLEMENTACIÓN DE UN PROGRAMA

Hay circunstancias que puedan comprometer una evaluación aleatorizada en el momento en el que una intervención está siendo implementada.

### 6.2.1 Es difícil mantener separados al grupo de comparación y al de tratamiento

Si es que, por criterios sociales o políticos, es difícil mantener separados al grupo de comparación y al grupo de intervención, la evaluación se verá amenazada. Una posible solución consiste en **asignar los grupos a diferentes proveedores del bien o servicio en cuestión**, para reducir la probabilidad de que las personas se cambien de grupo<sup>27</sup> (ver Figura 13). Sin embargo, si los miembros de ambos grupos son cercanos entre sí, pueden estar al tanto de la diferencia en proveedores y percibirlo como algo injusto. **Es importante que el tratamiento sea implementado de forma consistente entre los diferentes proveedores.**

---

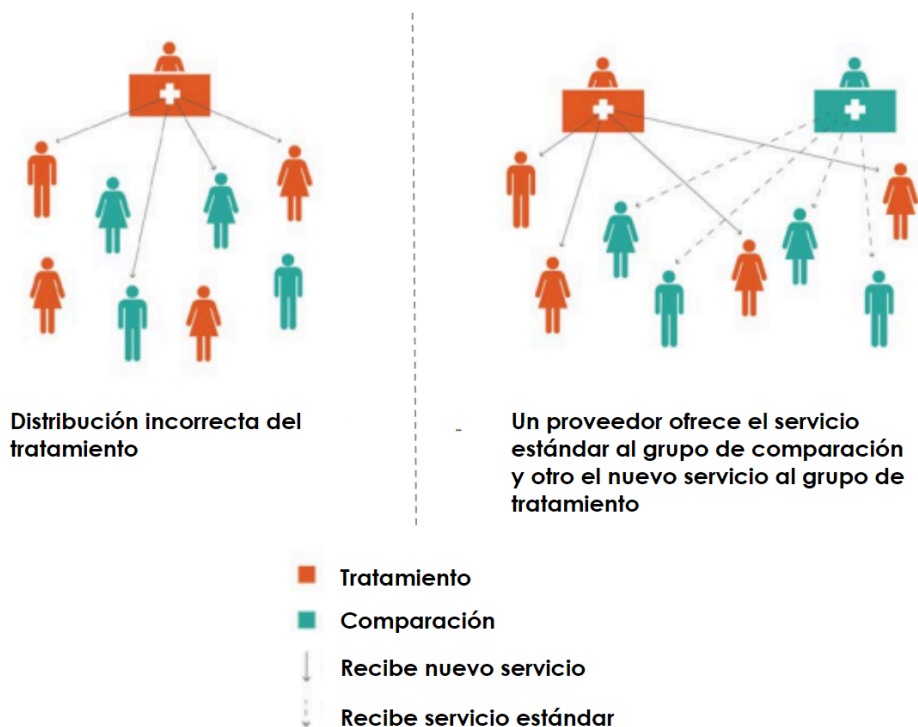
<sup>24</sup> Por ejemplo, si evaluamos un programa de construcción de plazas públicas y aleatorizar a nivel de hogar en vez de barrio, todas las personas del barrio tendrán acceso a la plaza sin importar si fueron asignados al grupo de tratamiento o comparación. En ese caso, aunque el programa fuera exitoso, concluiríamos que no tuvo un efecto.

<sup>25</sup> Como edad, nivel socioeconómico, ubicación, etc.

<sup>26</sup> Por ejemplo, si creemos que el programa es igualmente efectivo para mujeres y hombres, no resulta útil estratificar según sexo biológico.

<sup>27</sup> Por ejemplo, un proveedor ofrece una nueva capacitación al grupo de tratamiento y otro proveedor ofrece al grupo de comparación la capacitación habitual.

**Figura 13.** Uso de diferentes proveedores para evitar que el grupo de comparación acceda al nuevo servicio o tratamiento



Otra solución posible es **aumentar el nivel al que se aleatoriza**. Por ejemplo, en lugar de aleatorizar a nivel de hogar, aleatorizar a un nivel de barrio para que sea más fácil mantener separados a los miembros del grupo de comparación y de tratamiento. Una consecuencia de esta decisión es que se **reduce el número de unidades que pueden ser aleatorizadas** y, en consecuencia, el poder estadístico de la evaluación.

### 6.2.2 El grupo de comparación se entera del tratamiento, se beneficia del tratamiento o se ve afectado por el propio tratamiento

Los miembros del grupo de comparación podrían reaccionar negativamente si perciben que le están ofreciendo un servicio diferente, buscar beneficiarse del programa o sufrir efectos negativos como consecuencia de no participar en el programa<sup>28</sup>. Esto podría provocar que algunas personas se cambien de grupo o que se retiren del estudio.

Para prevenir lo anterior, se recomienda **aumentar el nivel al que se lleva a cabo la aleatorización**. Por ejemplo, se podría aleatorizar a nivel del vecindario en vez de a nivel de hogar, o a nivel de escuela en vez de a nivel de estudiante. La principal desventaja de aumentar el nivel de la aleatorización es que se pierde poder estadístico y por ende es más difícil saber si el programa fue efectivo o no.

<sup>28</sup> Por ejemplo, podría resultarles más difícil conseguir empleo si es que el grupo de tratamiento participa en una capacitación.

Otra opción es crear “amortiguadores”<sup>29</sup> que reduzcan la probabilidad de interacciones entre el grupo de tratamiento y el grupo de comparación. Los amortiguadores están conformados por unidades (personas, escuelas, firmas, etc.) que no forman parte del estudio, por lo que esta solución requiere que el proveedor del programa tenga la capacidad de cubrir un área extensa que incluye a los miembros del grupo de tratamiento, del grupo de comparación y de los amortiguadores.

**Figura 14.** Uso de “amortiguadores” para distanciar al grupo de tratamiento y comparación



**Recurso de profundización.** Esta sección se basa en el [manual](#) de J-PAL sobre algunos desafíos prácticos de las evaluaciones aleatorizadas (en inglés).

<sup>29</sup> “Buffer” en inglés.

## 7. UTILIZANDO LOS RESULTADOS DE EVALUACIONES EN POLÍTICA PÚBLICA

### 7.1 CAMINOS DE LA EVIDENCIA A LA ACCIÓN

La evidencia de las evaluaciones aleatorizadas está cambiando la forma en que entendemos y abordamos los problemas relacionados con la pobreza. Quienes formulan políticas, profesionales y organizaciones de todo el mundo están utilizando cada vez más las evaluaciones aleatorizadas para mejorar las políticas públicas y los programas sociales. La evidencia generada a partir de las evaluaciones aleatorizadas puede generar cambios a través de diferentes caminos.

El camino más obvio es que las evaluaciones aleatorizadas **entregan información clave para decidir si escalar, reducir o rediseñar el programa evaluado**. Un acercamiento hacia el diseño de políticas basadas en la evidencia es pilotear una innovación, evaluarla rigurosamente y luego escalar el piloto en el mismo contexto si este ha demostrado ser exitoso. En cambio, si los resultados no son los esperados, lo más conveniente es rediseñar el programa o incluso discontinuarlo.

En segundo lugar, la evidencia de que un programa funcionó en un contexto también puede llevarnos a **replicar el programa exitoso adaptándolo a un nuevo contexto**. Por ejemplo, varias evaluaciones aleatorizadas en India concluyeron que el programa educacional “Teaching at the Right Level (TaRL)”<sup>30</sup> mejoraba el aprendizaje de los estudiantes y fue adaptado y replicado en varios países de África.

Por otra parte, las evaluaciones aleatorizadas proveen **lecciones aplicables más allá** del programa evaluado. En vez de limitarse a determinar si un programa es efectivo o no, muchas evaluaciones aleatorizadas identifican los mecanismos detrás de su éxito, obteniendo lecciones generales que pueden ser aplicadas en otros contextos y sectores. Por ejemplo, las evaluaciones de TaRL sugieren que el mecanismo detrás del éxito del programa es enseñarle a cada estudiante lo que es capaz de aprender dados sus conocimientos actuales, en vez de enfocarse en lo que se supone que debería saber. Reconocer este mecanismo permite aplicarlo en programas aparentemente diferentes, pero que también se centran en adaptar la enseñanza al nivel del estudiante, como las tutorías o el aprendizaje asistido por computadora.

En cuarto lugar, los resultados de las evaluaciones aleatorizadas **pueden cambiar creencias y paradigmas**. En más de una ocasión, las evaluaciones aleatorizadas han desafiado el status quo, moldeando de manera fundamental nuestro entendimiento sobre un gran número de políticas sociales. Por ejemplo, por décadas se ha debatido la conveniencia de cobrar por productos básicos de salud preventiva, tales como redes mosquiteras, pastillas antiparasitarias o tratamientos de purificación de agua. Después de más de una docena de evaluaciones aleatorizadas en ocho países que sugieren que los costos de subsidiar los precios de estos productos superan los costos, varios gobiernos y organizaciones han optado por distribuir estos productos gratuitamente o a un precio rebajado.

---

<sup>30</sup> Ver más información en <https://www.povertyactionlab.org/case-study/teaching-right-level-improve-learning>.

Un último camino en que la evidencia cambia la política pública y los programas sociales es mediante la **institucionalización del uso evidencia**. Muchas organizaciones—incluyendo gobiernos e importantes organizaciones no-gubernamentales—han instalado unidades y generado procesos formales para diseñar y evaluar innovaciones de forma rigurosa e incorporar la evidencia en la toma de decisiones. Esta institucionalización puede tomar diferentes formas, desde la exigencia de incorporar literatura al diseño de programas hasta la creación de fondos de evaluación y la instalación laboratorios de innovación y evaluación.



**Recurso de profundización.** La sección “[Evidencia en acción](#)” de la web de J-PAL ofrece casos de estudios sobre cómo los resultados de evaluaciones aleatorizadas han informado el (re)diseño de programas y políticas.

## 7.2 APLICANDO EVIDENCIA A NUESTRO CONTEXTO

A lo largo del Ciclo de Aprendizaje mencionado al principio de este documento, existen varias instancias en las que es necesario utilizar evidencia: al diagnosticar el problema, al diseñar una nueva intervención y al momento de decidir qué hacer con el programa una vez que finalizamos la evaluación. En esta subsección se entregan algunas recomendaciones sobre cómo buscar evidencia, estimar su calidad, y aplicarla.

### 7.2.1 Buscando evidencia

En el contexto de esta guía, entenderemos evidencia como afirmaciones que se basan en datos. Por ejemplo, un podríamos afirmar que tenemos evidencia de que una persona está enferma, después de observar que tiene temperatura elevada (dato 1) y presenta sudoración (dato 2).

Una organización puede utilizar evidencia interna o externa. La evidencia interna es la que se genera dentro de la misma organización e incluye distintas fuentes como datos administrativos, resultados de monitoreo y evaluación de programas, datos de gestión y encuestas aplicadas por la organización, entre otros. La evidencia externa se genera fuera de la organización, como por ejemplo *papers* académicos, publicaciones de política pública, resultados de encuestas aplicadas por externos, estadísticas públicas, etc.

## RECUADRO 9: ALGUNAS FUENTES DE EVIDENCIA SUGERIDAS

Existen muchísimos lugares donde buscar información, lo importante es recurrir a fuentes confiables. A continuación, se dejan algunas sugerencias de dónde puede comenzar la búsqueda.

Para entender el contexto:

- Datos y reportes oficiales de los institutos de estadística del país
- Datos y reportes de instituciones confiables

Para conocer los resultados de evaluaciones individuales:

- [Resúmenes de evaluaciones de J-PAL](#)
- [3ie Evidence Gap Maps](#)
- [Hub de evaluación del BID](#)
- [Informes de eficacia de intervenciones - Givewell](#)

Para conocer evidencia agregada sobre qué intervenciones son efectivas:

- [Publicaciones de política pública de J-PAL](#)
- [J-PAL Policy Insights](#)
- [Campbell Systematic Reviews](#)

Otros recursos:

- [Reportes del Banco Mundial](#)
- Repositorios digitales como [JSTOR](#)
- Datos administrativos
- Datos de evaluación y monitoreo
- [J-PAL dataverse](#)
- [Google Scholar](#)

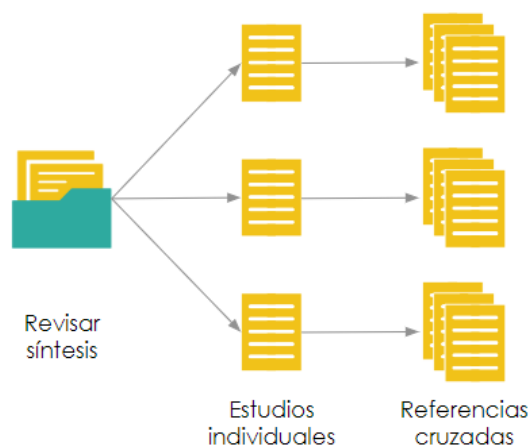
Al momento de buscar evidencia es importante definir la o las **preguntas que queremos responder, qué tipo de evidencia se incluirá o dejará fuera y cuál será nuestra ruta de búsqueda**. De esta forma nos aseguraremos de que nuestra revisión se realice en forma sistemática. Por ejemplo, podríamos definir como tema de interés los programas de crédito agrícola y decidir dejar fuera toda la evidencia que provenga de países de ingresos altos. Respecto a la ruta, existen varias opciones como seleccionar palabras clave para buscar publicaciones en sitios de búsqueda especializados como Google Scholar.

Otra ruta posible es la descrita en la Figura 15. El primer paso consiste en revisar una **síntesis de evidencia** como, por ejemplo, un metaanálisis o una revisión de literatura. Una vez que tengamos claridad sobre las principales conclusiones de la síntesis, podemos revisar individualmente los estudios incluidos en la síntesis que parecen más prometedores (por ejemplo, los que utilizan metodologías más robustas, los de contextos más similares al nuestro, o los que se refieren a programas parecidos al que estamos diseñando). Finalmente, cada uno de estos estudios individuales nos llevarán a otras

publicaciones<sup>31</sup>: las publicaciones que hacen referencia al estudio individual y las publicaciones que cita dicho estudio original.

Cabe preguntarse hasta qué punto debemos seguir recopilando evidencia. ¿Debemos incluir nuevas palabras clave? ¿Debemos revisar otras síntesis de evidencia? Si bien esta pregunta no tiene una respuesta clara, una señal de que la revisión ha sido amplia, es que al enfocarnos en una nueva palabra clave o síntesis, ya no nos encontramos con publicaciones individuales que no hayamos revisado.

**Figura 15.** Revisión de evidencia a partir de una síntesis



### 7.2.2 Estimando la calidad de la evidencia

La evidencia puede variar en su solidez dependiendo de los datos en que se basa. Volviendo a nuestro ejemplo anterior, la observación de que la persona tiene temperatura elevada será más confiable si la temperatura se mide con un termómetro, que si se mide con la palma de la mano. Asimismo, la calidad de la evidencia también está determinada por la rigurosidad del análisis que se hace a partir de los datos. Por ejemplo, la afirmación de que la persona está enferma probablemente esté equivocada si es que no consideró que la persona acababa de practicar deporte.

Estimar la calidad de la evidencia es un proceso complejo y que requiere de experiencia, no obstante, existen algunas preguntas que pueden orientar este proceso:

- ¿Cuál es la fuente de la evidencia?
  - ¿Síntesis de múltiples estudios, un solo estudio, reportes, blogs, periódicos, etc.?

---

<sup>31</sup> Aunque un estudio no es lo mismo que una publicación, por simplicidad en este párrafo asumimos que cada publicación corresponde a un estudio y utilizamos ambos términos indistintamente.

- Si es una síntesis de varios estudios:
  - ¿Cuáles son los criterios de búsqueda e inclusión (cómo se decidió que estudios se incluyen en la síntesis)?
  - ¿Son los criterios de comparación claros y apropiados?
  - ¿Cómo se hace cargo la revisión de la diferencia en la calidad de los estudios individuales?
- Si es solo un estudio:
  - ¿Lo revisaron pares?
  - ¿Cuál es el diseño de la investigación? ¿Descriptivo, cuasi-experimental, aleatorizada, etc.?
  - ¿Qué supuestos se hicieron? ¿Se cumplen?
  - Para afirmaciones causales, ¿existe un grupo de comparación válido y/o una teoría del cambio?
  - ¿Cuál es el tamaño de la muestra, cómo se llevó a cabo la implementación, tasa de adopción, etc.?

### 7.2.3 Determinando la relevancia de la evidencia en nuestro contexto

Contar con evidencia robusta no necesariamente implica que esa evidencia es relevante para el contexto en que se quiere aplicar. ¿Cómo saber entonces si lo aprendido a partir de evaluaciones en otros lugares es relevante para nuestro problema?

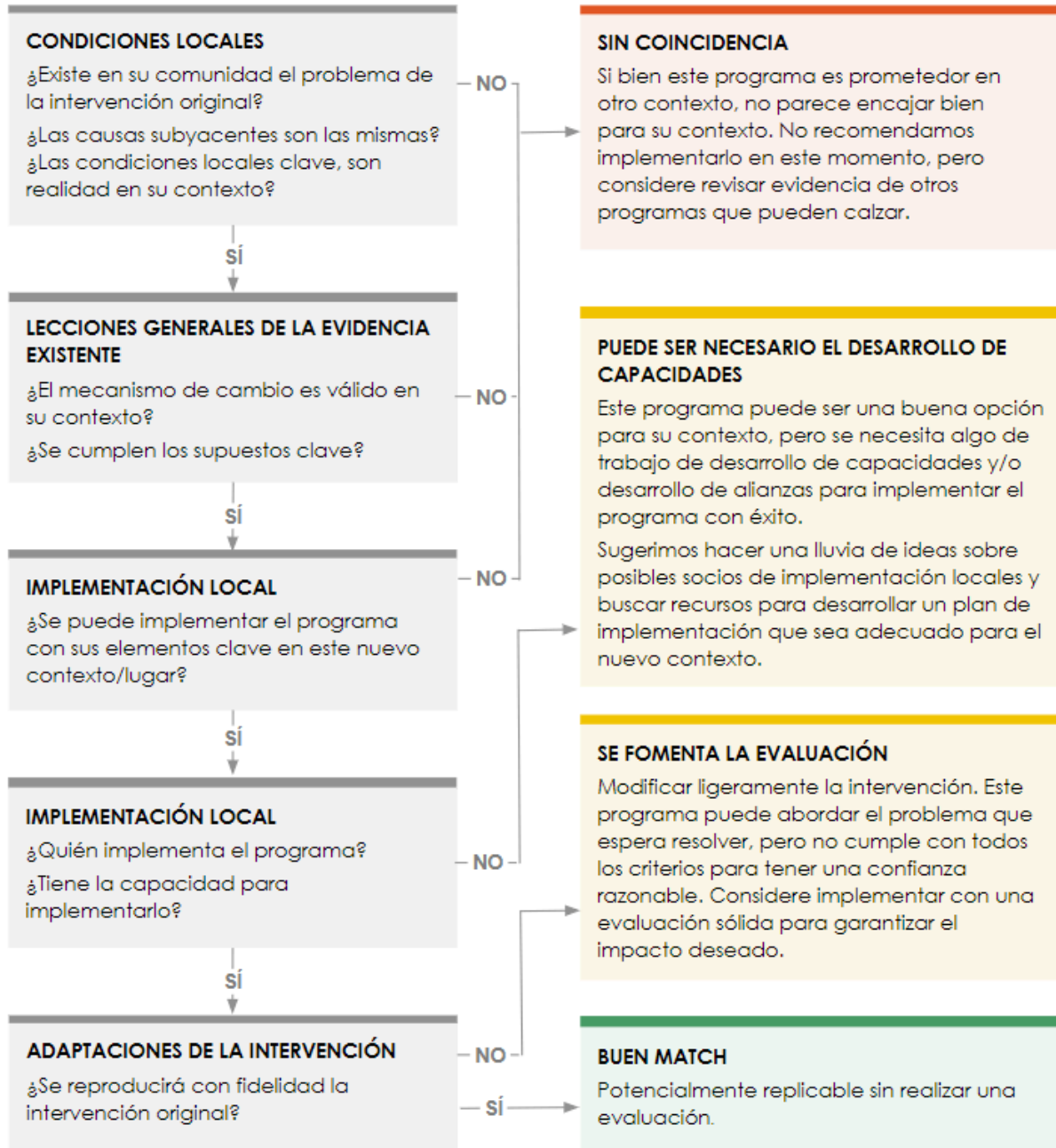
El **Marco de Generalizabilidad**<sup>32</sup> es una herramienta teórica que permite identificar cuáles son las evidencias más relevantes para nuestro contexto. El Marco de Generalizabilidad **se centra en los mecanismos causales, y no en características específicas de un programa**, lo que permite generalizar aprendizajes y aplicarlos en otros contextos. Se compone de cuatro etapas:

1. **Comprender la Teoría de Cambio.** En lugar de replicar un programa de forma idéntica, es necesario centrar la atención en el mecanismo causal que hizo eficaz al programa.
2. **Analizar las condiciones locales.** Encontrar datos descriptivos para entender mejor si el problema subyacente del contexto original también está presente en su comunidad.
3. **Analizar la robustez de la evidencia** de los mecanismos causales que hicieron al programa eficaz y si sus principales supuestos se mantienen en el contexto local.
4. **Evaluar si su organización** (u otra) puede implementar exitosamente la intervención, en apego al modelo original.

La Figura 16 esquematiza cómo utilizar el Marco de Generalizabilidad para aplicar evidencia en nuevos contextos.

<sup>32</sup> Ver Bates, M. A., and Glennerster, R. 2017. “The Generalizability Puzzle”. *Stanford Social Innovation Review*, 15(3), 50–54. <https://doi.org/10.48558/EYY5-3S89>

**Figura 16.** Flujo de decisiones al evaluar la aplicabilidad de la evidencia a un nuevo contexto



## RECUADRO 10: MARCO DE GENERALIZABILIDAD EN UN PROGRAMA DE TUTORÍAS ESCOLARES

Supongamos que un gobierno quiere mejorar el aprendizaje de los estudiantes en matemáticas. Después de que una evaluación aleatorizada encontrara que un programa de tutorías en Estados Unidos aumentó los puntajes de los estudiantes en pruebas estandarizadas, un gobierno latinoamericano está considerando implementar un programa de ese tipo. ¿Cómo saber si los resultados de esa evaluación son relevantes en este nuevo contexto?

1. **Comprender la Teoría de Cambio.** En primer lugar, es necesario conocer cabalmente la Teoría de Cambio del programa para entender cuál es el mecanismo que causó el impacto. Tal como indica Teoría de Cambio del Recuadro 2, el mecanismo que facilita el aprendizaje de los estudiantes es que se personaliza el nivel de enseñanza a lo que cada estudiante sabe.
2. **Analizar las condiciones locales.** Debemos analizar si se dan las condiciones locales para que la Teoría de Cambio se cumpla en este contexto. Por ejemplo, el programa de tutorías se basa en el supuesto de que los estudiantes asisten a la escuela, por lo que si los problemas de aprendizaje se deben a inasistencia, es poco probable que las tutorías sean efectivas.
3. **Analizar la robustez de la evidencia.** Existe un cuerpo importante de evidencia que indica que las tutorías son efectivas para aumentar los puntajes de los estudiantes en matemáticas. Además, también existe evidencia prometedora de otros programas que se basan en el mecanismo de adaptar el nivel de enseñanza a los conocimientos de cada estudiante, como el aprendizaje asistido por computadora o agrupar a los estudiantes dentro de la sala de clase según sus conocimientos.
4. **Evaluar las capacidades locales.** Se debe indagar en la capacidad del gobierno de replicar el mecanismo de cambio. ¿Cuenta con el personal necesario para hacerlo? ¿Permiten el currículum y la normativa realizar las tutorías? ¿En qué horario se harán?

## 7.3 EXPANDIENDO UN PROGRAMA

### 7.3.1 Eligiendo entre varios programas: costo-efectividad

Para poder invertir los recursos óptimamente, es importante tomar en cuenta los costos de un programa. La costo-efectividad reúne en un único indicador el impacto y los costos, indicando el efecto de un programa por cada dólar (u otra unidad monetaria) invertido. Así, es posible comparar programas evaluados en diferentes contextos. Alternativamente, podemos entender la costo-efectividad como el costo que tiene lograr un impacto deseado<sup>33</sup>.

Un análisis de costo efectividad se puede hacer en forma **retrospectiva** (calcular los costos y efectos del programa después de su implementación) o **prospectiva** (anticipar los costos y efectos del programa antes de su implementación con base en programas similares).

<sup>33</sup> Por ejemplo, cuánto cuesta aumentar la cantidad de trabajadores de una empresa en 0,1 trabajadores en promedio.

### 7.3.2 Analizando los cambios en efectividad y costos al escalar un programa

Aunque una evaluación indique que un programa es costo-efectivo, es necesario hacer un análisis cuidadoso antes de expandirlo, considerando cómo pueden variar la efectividad y los costos del programa al cambiar la escala. El análisis dependerá de cada contexto, pero a continuación se presentan algunas preguntas guía.

#### Cambios en la efectividad

- ¿La población participante en el programa es similar a la población que participó en la evaluación?
- ¿Quién implementará el programa a mayor escala? ¿Tiene las capacidades para mantener la implementación tal como se hizo en el programa original?
- ¿Existen posibles mejoras para aumentar la efectividad del programa?

#### Cambios en los costos

- ¿Implicará la expansión del programa la necesidad de inversiones que no se hicieron durante el programa original?
- ¿Es posible utilizar los insumos de la misma forma?
- ¿Existen economías de escala (es decir, el costo por participante del programa disminuye a medida que más personas participan)?
- ¿Existen costos fijos independientes de la cantidad de personas que participen? (ej.: desarrollo de materiales)

#### Otras consideraciones

- ¿Están disponibles los insumos para escalar el programa? Por ejemplo, si el programa necesita de personal altamente cualificado, a medida que lo escalamos será más difícil encontrar personas que cumplan con el perfil adecuado.

## 8. CASO DE ESTUDIO: EVALUACIÓN ALEATORIZADA DE PRINCIPIO A FIN

En Chile, la sobrepesca de la merluza del Pacífico representa un gran desafío ambiental y económico. Esta práctica puede dañar ecosistemas marinos, crear tensiones económicas en los mercados, y amenazar las necesidades nutricionales y de consumo de las poblaciones locales. El investigador J-PAL, Mushfiq Mobarak (Universidad de Yale) y su coautor Andrés González-Lira (UC Berkeley), en conjunto al Servicio Nacional de Pesca (Sernapesca) evaluaron diferentes estrategias para asegurar el cumplimiento de la veda estacional para la merluza<sup>34</sup>. Su evaluación aleatorizada encontró que tanto las campañas de información a consumidores como las estrategias de fiscalización de las ventas fueron efectivas en reducir la disponibilidad de merluza ilegal en los mercados locales. Guiados por los resultados de la evaluación y un análisis de costo-efectividad, Sernapesca expandió la campaña de conciencia al consumidor evaluada en el estudio y adaptó sus tácticas de fiscalización.

### El problema

A pesar del esfuerzo del gobierno, muchas especies marinas en Chile han sido sobreexplotadas y están en riesgo de colapsar. El gobierno chileno ha creado numerosas regulaciones para proteger especies en riesgo, incluyendo la implementación de cuotas restrictivas y veda a recolección y venta de peces durante las épocas de reproducción. Sin embargo, muchas especies de peces continúan en riesgo.

La implementación y fiscalización de las cuotas de pesca han resultado difíciles debido a los costos programáticos restrictivos. La mayor parte de la pesca en Chile es realizada a pequeña escala por personas que operan de manera informal en negocios geográficamente dispersos, dificultando el monitoreo. Un desafío adicional que enfrenta el gobierno es la habilidad de pescadores y vendedores para encontrar vacíos y evadir al personal que regula, bajando la efectividad de las actividades regulatorias y permitiendo el florecimiento de los negocios ilícitos. Además, hay una falta general de conciencia entre quienes consumen pescado sobre las regulaciones gubernamentales en la industria pesquera.

### La investigación

En mayo del 2015, oficiales del Servicio Nacional de Pesca asistieron a un curso de incubación de proyectos dirigido por la oficina de J-PAL Latinoamérica y Caribe y la Dirección de Presupuestos de Chile (DIPRES). Luego del curso, algunas de las personas asistentes se interesaron en desarrollar y evaluar rápidamente diversas estrategias regulatorias. En conjunto a investigadores asociados a J-PAL testearon los impactos relativos y la costo-efectividad de las campañas de información orientadas al consumidor y de la fiscalización de los puestos de venta sobre la captura y comercio ilegales de merluza.

Los investigadores asignaron aleatoriamente a conjuntos de vendedores (también conocidos como circuitos) a uno de cuatro grupos:

---

<sup>34</sup> Gonzalez Lira, Andres, and Ahmed Mushfiq Mobarak. Slippery fish: Enforcing regulation under subversive adaptation. Working Paper, March 2021. <https://www.nber.org/papers/w28610>

1. Campaña de información: el Servicio Nacional de Pesca distribuyó cartas, panfletos y posters en los sectores residenciales alrededor del circuito. Los materiales informaban al consumidor sobre la disminución de la población de la merluza y la veda en las ventas.
2. Fiscalización: agentes inspectores realizaron visitas regulares a los mercados de pecados del circuito, entregando sanciones de US\$200 a quienes eran observados vendiendo pescado ilegalmente.
3. Información y fiscalización: los circuitos recibieron la campaña de información y la fiscalización.
4. Grupo de comparación: los circuitos no recibieron ni la campaña de información ni la fiscalización.

Los grupos que recibieron el tratamiento de fiscalización (2 y 3) fueron aleatoriamente asignados a dos condiciones más: recibir visitas predecibles o impredecibles y recibir visitas de baja o alta intensidad.

Las campañas de información y fiscalización redujeron la disponibilidad y consumo de la merluza en un 30 por ciento. La campaña de información llevó a más consumidores a estar alerta de la veda y reportó un menor consumo ilegal de merluza. Las fiscalizaciones que ocurrieron en forma impredecible llevaron a reducciones sustanciales de merluza ilegal en los mercados locales. Sin embargo, las fiscalizaciones más frecuentes con aviso previo no redujeron las ventas ilegales. En su lugar, las y los vendedores adoptaron tácticas para evitar las sanciones, como esconder el pescado o decir que había sido capturado y congelado antes de la veda.

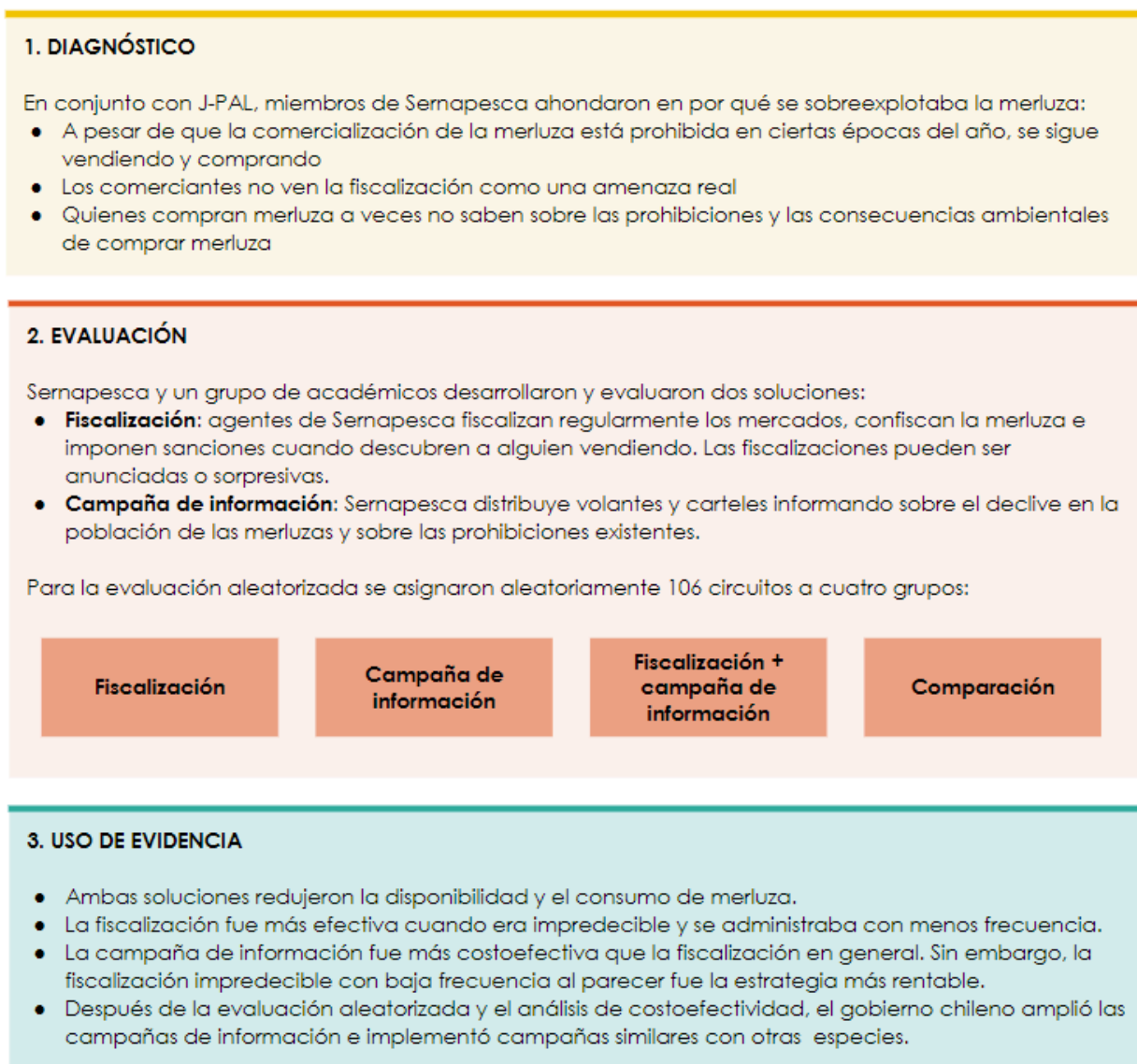
Los investigadores también recolectaron información de Sernapesca sobre los costos administrativos de implementar ambas estrategias para realizar un análisis comparativo de costo-efectividad. Encontraron que la información de la campaña era más costo-efectiva en reducir la venta ilegal de merluza que la estrategia de fiscalización.

### De la investigación a la acción

La evaluación aleatorizada convenció al Servicio Nacional de Pesca de la efectividad de las campañas de información, las cuales continuaron y expandieron en los años subsecuentes. Orientada por los resultados de la evaluación y el análisis de costo-efectividad, la agencia ha continuado desde el 2016 la campaña de información durante la veda estacional a la pesca de merluza. El Servicio Nacional de Pesca también cambió la forma en que realizaban las fiscalizaciones. Usando como base los hallazgos y datos recolectados de la evaluación, adaptó su estrategia de fiscalización llegar a los mercados más estratégicos de forma impredecible.

Sernapesca ha aplicado los hallazgos de las evaluaciones para diseñar campañas similares para otras especies sobreexplotadas. En el 2018, la agencia lanzó un nuevo sitio web y el hashtag #salvavedas para promover la protección de doce especies marinas sobreexplotadas. La agencia también expandió la estrategia de sus campañas de información, colgando *posters*, distribuyendo panfletos y lanzando campañas locales y en redes sociales para seis especies marinas.

**Figura 17.** Resumen del Ciclo de Aprendizaje en el programa de Sernapesca



## REFERENCIAS

A continuación se detallan los documentos en los que se basaron cada una de las secciones:

### 1. Introducción

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2018. “Forjando una Cultura para el Uso de Evidencia: Lecciones de J-PAL sobre sus Alianzas con Gobiernos en Latinoamérica”.

<https://www.povertyactionlab.org/sites/default/files/creating-a-culture-of-evidence-use-lessons-from-jpal-govt-partnerships-in-latin-america-spanish.pdf>

International Labour Organization. 2015. “Basic Principles of Monitoring and Evaluation”.

[https://www.ilo.org/employment/areas/youth-employment/WCMS\\_546505/lang-en/index.htm](https://www.ilo.org/employment/areas/youth-employment/WCMS_546505/lang-en/index.htm)

### 2. Entendiendo cómo queremos generar cambios y medir impacto

Bernhardt, Arielle, Erica Field, Rohini Pande, and Natalia Rigol. 2019. "Household matters: Revisiting the returns to capital among female microentrepreneurs." *American Economic Review: Insights* 1, no. 2: 141-160. DOI: 10.1257/aeri.20180444

Glennster, R. and Takavarasha, K. 2014. “Running randomized evaluations: A practical guide”. *Princeton University Press*.

J-PAL. 2023. “Lecture: Theory of Change and Measurement.” *Abdul Latif Jameel Poverty Action Lab, Cambridge, MA*.

[https://www.povertyactionlab.org/sites/default/files/research-resources/ToCandMeasurement\\_Lecture\\_Slides\\_2023.pdf](https://www.povertyactionlab.org/sites/default/files/research-resources/ToCandMeasurement_Lecture_Slides_2023.pdf)

### 3. Evaluando impacto

Gertler, Paul J., Sebastián Martínez, Patrick Premand, and Laura B. Rawlings. 2017. “La evaluación de impacto en la práctica”. *World Bank Publications*. Capítulos 1 y 3.

<https://openknowledge.worldbank.org/server/api/core/bitstreams/6f2eebf7-1a3c-5f67-a9c3-c39f68299ed9/content>

Gibson, Michael and Anja Sautmann. “Introduction to randomized evaluations”. *Abdul Latif Jameel Poverty Action Lab, Cambridge, MA*.

<https://www.povertyactionlab.org/resource/introduction-randomized-evaluations>

Glennster, R. and Takavarasha, K., 2014. “Running randomized evaluations: A practical guide.” *Princeton University Press*. Capítulo 2.

J-PAL. 2023. “Lecture: Why Evaluate.” *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*  
[https://www.povertyactionlab.org/sites/default/files/research-resources/WhyEvaluate\\_LectureSlides\\_2023.pdf](https://www.povertyactionlab.org/sites/default/files/research-resources/WhyEvaluate_LectureSlides_2023.pdf)

#### **4. Las evaluaciones de impacto aleatorizadas**

Gertler, Paul J., Sebastián Martínez, Patrick Premand, and Laura B. Rawlings. 2017. “La evaluación de impacto en la práctica”. *World Bank Publications*. Capítulo 4.  
<https://openknowledge.worldbank.org/server/api/core/bitstreams/6f2eebf7-1a3c-5f67-a9c3-c39f68299ed9/content>

Glennester, R. and Takavarasha, K., 2014. “Running randomized evaluations: A practical guide.” *Princeton University Press*. Capítulo 4.

J-PAL. 2023. “Lecture: Why Randomize.” *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA*  
[https://www.povertyactionlab.org/sites/default/files/research-resources/WhyandWhentoRandomize\\_LectureSlides\\_2023.pdf](https://www.povertyactionlab.org/sites/default/files/research-resources/WhyandWhentoRandomize_LectureSlides_2023.pdf)

#### **5. Consideraciones éticas al implementar evaluaciones**

Comisión Nacional para la Protección de Sujetos Humanos de Investigación Biomédica y de Comportamiento. 1979. “Informe Belmont: Principios Éticos y Directrices para la Protección de Sujetos Humanos de Investigación”. <https://www.hhs.gov/sites/default/files/informe-belmont-spanish.pdf>

Feeney, Laura, Sarah Kopper, and Anja Sautmann. 2022. “Ethical conduct of randomized evaluations”. *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*  
<https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations>

J-PAL. 2023. “Lecture: Ethics of Conducting Randomized Evaluations”. *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*  
[https://www.povertyactionlab.org/sites/default/files/EthicalConsiderations\\_LectureSlides\\_2023.pdf](https://www.povertyactionlab.org/sites/default/files/EthicalConsiderations_LectureSlides_2023.pdf)

#### **6. Desafíos prácticos en una evaluación aleatorizada**

Feeney, Laura, Sarah Kopper, and Anja Sautmann. 2022. “Ethical conduct of randomized evaluations”. *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*  
<https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations>

Heard, Kenya, Elisabeth O’Toole, Rohit Naimpally, and Lindsey Bressler. 2017. “Real world challenges to randomization and their solutions”. *Boston, MA: Abdul Latif Jameel Poverty Action Lab.*

<https://www.povertyactionlab.org/sites/default/files/research-resources/2017.04.14-Real-World-Challenges-to-Randomization-and-Their-Solutions.pdf>

## 7. Utilizando los resultados de evaluaciones en política pública

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2023. “Lecture: The Generalizability Framework.” *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*

[https://www.povertyactionlab.org/sites/default/files/Generalizability\\_LectureSlides\\_2023.pdf](https://www.povertyactionlab.org/sites/default/files/Generalizability_LectureSlides_2023.pdf)

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2022. “Teaching at the Right Level to improve learning”.

<https://www.povertyactionlab.org/case-study/teaching-right-level-improve-learning>

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2020. “The Transformative Potential of Tutoring for Pre K-12 Learning Outcomes: Lessons from Randomized Evaluations”. *Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.*

<https://www.povertyactionlab.org/publication/transformative-potential-tutoring-pre-k-12-learning-outcomes-lessons-randomized>

Bates, M. A., and Glennerster, R. 2017. “The Generalizability Puzzle”. *Stanford Social Innovation Review*, 15(3), 50–54. <https://doi.org/10.48558/EYY5-3S89>

## 8. Caso de estudio: evaluación aleatorizada de principio a fin

Gonzalez Lira, Andres, and Ahmed Mushfiq Mobarak. “Slippery fish: Enforcing regulation under subversive adaptation”. *Working Paper, March 2021.* <https://www.nber.org/papers/w28610>

## GLOSARIO

**Contrafactual.** Al momento de evaluar un programa, se refiere al escenario hipotético de lo que hubiera sucedido en ausencia del programa.

**Costo-efectividad:** Comparación del costo relativo de diferentes alternativas en función de su capacidad para alcanzar un objetivo específico. *Ejemplo* - Comparar el costo y el número de enfermedades prevenidas por dos diferentes campañas de vacunación.

**Datos:** Conjunto de información cuantitativa o cualitativa recolectada, utilizada como base para análisis y decisiones. *Ejemplo* - Recopilación de cifras de desempleo y niveles de educación en diferentes regiones para diseñar políticas públicas.

**Efectividad:** Grado en que una actividad o intervención cumple sus objetivos en condiciones reales de implementación. *Ejemplo* - Medir el porcentaje de reducción de la delincuencia tras implementar un nuevo programa de vigilancia policial.

**Evidencia:** Datos e información obtenidos mediante métodos sistemáticos, usados para fundamentar conclusiones o decisiones. *Ejemplo* - Utilizar estudios científicos para justificar la implementación de un nuevo sistema de transporte público.

**Factual.** Al momento de evaluar un programa, se refiere a lo que sucedió una vez implementado el programa.

**Marco de Generalizabilidad:** Una herramienta que identifica evidencias relevantes para aplicar aprendizajes de un contexto a otro, enfocándose en mecanismos causales y condiciones locales. *Ejemplo* - Si un programa de educación mejoró el rendimiento escolar en una región, el Marco de Generalizabilidad puede ayudar a comprender si el mismo enfoque podría ser efectivo en otra región con características similares.

**Marco Lógico:** Herramienta para planificar y gestionar proyectos, definiendo objetivos, actividades, resultados esperados y medios de verificación. *Ejemplo* - Establecer los objetivos, actividades, y resultados esperados para un proyecto de construcción de infraestructura escolar.

**Metaanálisis:** Método estadístico que integra/combina resultados de varios estudios para obtener una conclusión global sobre un tema específico. *Ejemplo* - Combinar resultados de estudios sobre la eficacia de programas de alimentación escolar para obtener una visión global.

**Poder.** En este contexto, el poder estadístico se refiere a la probabilidad de que la evaluación detecte el impacto de un programa efectivo. *Ejemplo* - Si un programa efectivamente tiene un impacto positivo, una evaluación con poder estadístico del 80 por ciento concluirá que el programa es efectivo con 80 por ciento de probabilidad.

**Revisión de literatura:** Proceso de recopilar, analizar y sintetizar investigaciones existentes sobre un tema para obtener una comprensión integral. *Ejemplo* - Analizar diversas investigaciones sobre políticas de vivienda asequible para desarrollar un nuevo marco normativo.