

Ensayos controlados aleatorios

Howard White, Shagun Sabarwal y Thomas de Hoop

El Centro de Investigaciones Innocenti es la oficina de UNICEF especializada en investigación científica. El objetivo principal del Centro de Investigaciones es mejorar la comprensión internacional de una serie de cuestiones relacionadas con los derechos de la infancia, a fin de facilitar la plena aplicación de la Convención sobre los Derechos del Niño en todo el mundo. El centro tiene el cometido de establecer un marco integral de investigación y conocimiento dentro de la organización para brindar apoyo a los programas y políticas mundiales de UNICEF, y trabaja con los asociados para formular políticas con base empírica en favor de la infancia. Las publicaciones elaboradas por el centro contribuyen al debate global sobre la infancia y los derechos del niño e incluyen una amplia gama de opiniones.

Las opiniones expresadas corresponden a los autores o editores y se publican para estimular un mayor diálogo sobre métodos de análisis de impacto. Esta publicación no refleja necesariamente las políticas o perspectivas de UNICEF.

SINTESIS METODOLOGICAS DEL CENTRO DE INVESTIGACIONES

Las síntesis metodológicas del Centro de Investigaciones de UNICEF pretenden compartir prácticas de investigación, métodos, diseños y recomendaciones de reconocidos investigadores y analistas. La audiencia a la que van dirigidas es principalmente el personal de UNICEF que lleve a cabo, encargue o interprete los resultados de investigación y análisis para la toma de decisiones sobre programas, políticas y actividades de sensibilización.

Esta síntesis metodológica ha seguido un proceso de revisión por pares interna.

El texto no ha sido editado de acuerdo con los estándares de publicación oficiales y UNICEF declina toda responsabilidad por posibles errores.

Se permite la reproducción de cualquier parte de la presente publicación siempre que se incluya referencia a la presente. Si se desea utilizar una parte sustancial o la totalidad de la publicación dirijan su solicitud al Departamento de Comunicación en la dirección de correo electrónico: florence@unicef.org

Para consultas o descargas, pueden encontrar estas síntesis metodológicas en <http://www.unicef-irc.org/KM/IE/>

Recomendamos la siguiente cita para cualquier referencia al presente documento:

White, H., Sabarwal S. y T. de Hoop, (2014). Ensayos controlados aleatorios, *Síntesis metodológicas: evaluación de impacto n.º 7*, Centro de Investigaciones de UNICEF, Florencia.

Agradecimientos: Varios autores han proporcionado orientación en la preparación de esta síntesis. El autor y el Centro de Investigaciones de UNICEF desean agradecer a todos aquellos que han participado en la preparación de la presente publicación, especialmente a:

Por su contribución: Dugan Fraser, Simon Hearn, Patricia Rogers, Jessica Sinclair Taylor

Por su revisión: Nikola Balvin, Samuel Bickel, Sarah Hague, Sudhanshu Handa, Aisha Yousafzai

© Fondo de las Naciones Unidas para la Infancia (UNICEF), septiembre de 2014

Centro de Investigaciones Innocenti de UNICEF

Piazza SS. Annunziata, 12

50122 Florencia (Italia)

Tel.: (+39) 055 20 330

Fax: (+39) 055 2033 220

florence@unicef.org

www.unicef-irc.org

1. ENSAYOS CONTROLADOS ALEATORIOS: BREVE DESCRIPCIÓN

El ensayo controlado aleatorio es un método de evaluación de impacto en el que la población beneficiaria de la intervención del programa o la política y el grupo de control se eligen de manera aleatoria entre la población que cumple los criterios. Evalúa en qué medida se están alcanzando los impactos específicos planeados.

En el ensayo controlado aleatorio, el programa o la política se consideran una «intervención» cuyo tratamiento —los elementos del programa o la política objeto de evaluación— se pone a prueba a fin de determinar en qué medida logra sus objetivos, en función de un conjunto de indicadores predeterminado. Uno de sus puntos fuertes es que proporciona una respuesta contundente a las preguntas de causalidad y ayuda a los evaluadores y ejecutores del programa a determinar si lo que se está consiguiendo se debe únicamente a la intervención.

El ensayo controlado aleatorio es una forma [experimental](#) de evaluación de impacto; las formas cuasiexperimentales y no experimentales de evaluación de impacto se tratan en otros capítulos de esta serie: en la Síntesis n.º 6 (Sinopsis: Estrategias de atribución causal), la Síntesis n.º 8 (Diseños y métodos cuasiexperimentales) y la Síntesis n.º 9 (Estudios de caso comparativos).

La característica distintiva de los ensayos controlados aleatorios es la distribución aleatoria de los miembros de un grupo de población idónea entre uno o varios [grupos de tratamiento](#) (que reciben el tratamiento de la intervención¹ o sus variantes) y el grupo de control (que, o bien no recibe ninguna intervención, o bien recibe la intervención habitual, si el tratamiento objeto de estudio es una versión revisada de otro anterior). A continuación, se comparan los efectos en ámbitos de impacto específicos entre los grupos tras un período establecido de antemano. El cuadro 1 resume la diferencia entre la distribución aleatoria y el muestreo aleatorio, dos elementos clave de los ensayos controlados aleatorios.

Cuadro 1. Distribución aleatoria frente a muestreo aleatorio

No debe confundirse la distribución aleatoria con el muestreo aleatorio. El muestreo aleatorio se refiere a la toma de una muestra de uno o varios grupos de población. La distribución aleatoria consiste en la distribución de individuos o grupos entre los grupos de tratamiento y el grupo de control. Normalmente, los ensayos controlados aleatorios utilizan tanto el muestreo aleatorio (ya que por lo general tratan de obtener inferencias sobre una población mayor) como la distribución aleatoria (elemento característico del ensayo controlado aleatorio).

El diseño de ensayo controlado aleatorio más sencillo consta de un grupo de tratamiento y un grupo de control. Las variantes del diseño pueden contar con:

- **varios grupos de tratamiento**; por ejemplo, uno de los grupos recibe la intervención A y el segundo grupo recibe la intervención B, o
- **un diseño factorial**, en el que un tercer grupo de tratamiento recibe las intervenciones A y B.

¹ Pueden emplearse ensayos controlados aleatorios para medir tanto intervenciones programáticas (por ejemplo, la distribución de suplementos nutricionales en el marco de un programa sobre nutrición) como intervenciones sobre políticas (por ejemplo, la distribución de efectivo a raíz de una política de transferencia de efectivo). En aras de la brevedad, en esta síntesis sencillamente denominaremos «programa» o «intervención» a cualquiera de esos tipos de intervención.

Cuando ya hay una intervención en curso, es preferible que el grupo de control siga recibéndola, y que el ensayo controlado aleatorio compare la nueva intervención con la primera.

En un ensayo controlado aleatorio simple, se aplica la misma unidad de análisis a la intervención y a la distribución aleatoria. Por ejemplo, si se evalúa un programa que facilita alimentos a un conjunto de individuos, estos pueden distribuirse aleatoriamente para recibir suplementos alimenticios.

Sin embargo, por razones tanto prácticas como éticas, lo habitual es emplear un [diseño por conglomerados](#), en el que la unidad de distribución contiene múltiples unidades de tratamiento. Por ejemplo, las intervenciones educativas suelen distribuirse por escuelas, aunque la intervención tiene lugar entre los profesores, en las clases o con cada niño, y sus efectos se miden en los niños. Las intervenciones nutricionales, por otra parte, pueden distribuirse a nivel comunitario o de subdistritos. Dado el tipo de programas a gran escala que respalda UNICEF, lo más probable es que se recurra a ensayos controlados aleatorios por conglomerados.

Puntos principales

1. Un ensayo controlado aleatorio mide el efecto de una intervención de un programa o política en un resultado concreto.
2. Su característica principal es el uso de la distribución **aleatoria** de una intervención. Este diseño se denomina «diseño experimental»
3. Los ensayos controlados aleatorios solo sirven para medir impactos en determinadas situaciones, entre otras cuando es posible acceder a una muestra de tamaño amplio; cuando los efectos previstos de la intervención del programa o la política se pueden definir y medir fácilmente (como la reducción del retraso en el crecimiento); y cuando los ensayos controlados aleatorios se planifican antes del comienzo de la intervención.

2. ¿CUÁNDO PROCEDE EMPLEAR ESTE MÉTODO?

Un ensayo controlado aleatorio debe planificarse antes del inicio del programa

Un ensayo controlado aleatorio debe planificarse antes de la puesta en marcha del programa, y la participación en este ha de controlarse minuciosamente con miras al experimento. Los ensayos controlados aleatorios no pueden llevarse a cabo de forma retrospectiva.

Un caso excepcional lo encontramos en el [diseño basado en el estímulo](#), que no asigna aleatoriamente a los participantes a una intervención propiamente dicha, sino que les hace llegar material promocional o información adicional sobre los beneficios de una intervención a su alcance, a fin de fomentar su participación. Pueden emplearse diseños basados en el estímulo cuando un programa universalmente disponible no ha sido adoptado por todos.

Los ensayos controlados aleatorios necesitan una muestra amplia

Un ensayo controlado aleatorio solo es viable cuando el tamaño de la muestra es lo bastante amplio para detectar los efectos del programa con suficiente precisión; el diseño del estudio debe contar con lo que los estadísticos denominan «potencia estadística» suficiente.

Se trata de la probabilidad de llegar a conclusiones acertadas sobre la eficacia de un programa. En el proceso de diseño de un ensayo controlado aleatorio se llevan a cabo cálculos de potencia a fin de determinar qué tamaño debe tener la muestra para detectar el impacto del programa (véase el cuadro 2). Cuanto mayor es la muestra, mayor es su potencia.

En los ensayos controlados aleatorios por conglomerados, es el número de grupos —no el número de observaciones— el que determina la potencia estadística del estudio. Por ejemplo, una muestra de 50 comunidades en la que se toman 5 hogares por comunidad tiene mucha más potencia (50 grupos) que una en la que participen 25 comunidades de las que se tomen 10 hogares (25 grupos), a pesar de que el tamaño total de la muestra sea en ambos casos de 250 hogares.

Existen programas informáticos capaces de ejecutar cálculos de potencia (por ejemplo, Optimal Design), pero es preferible que esta tarea la desempeñe una persona con experiencia.

Cuadro 2. Cálculos de potencia estadística

La potencia estadística se refiere a la probabilidad de detectar un impacto cuando un programa lo propicia. Para efectuar cálculos de potencia y calibrar el tamaño de muestra que requiere una evaluación, los evaluadores suelen aplicar una serie de supuestos sobre la magnitud del efecto previsto, la significación estadística y la correlación en los conglomerados (en el caso de los ensayos controlados aleatorios por conglomerados). La correlación en los conglomerados es una estadística descriptiva de 0 a 1 que indica el grado de semejanza entre los grupos (por ejemplo, los hogares) o los individuos de un conglomerado. Cuanto mayor sea esa correlación, mayor habrá de ser el tamaño de la muestra. En los ensayos controlados aleatorios por conglomerados, normalmente la potencia estadística aumenta en mayor medida cuando se incrementa el número de conglomerados que cuando aumenta el número de individuos o grupos dentro de un conglomerado.

Los ensayos controlados aleatorios deben llevarse a cabo después de una investigación o evaluación formativa

El empleo de un ensayo controlado aleatorio para evaluar un programa que todavía no ha alcanzado la madurez probablemente resultará inadecuado. Además, la mayoría de las veces no debe efectuarse hasta que el programa se desarrolle de manera apropiada. Cabe trazar un paralelismo con el proceso de los ensayos clínicos con medicamentos, que contemplan un período de desarrollo y prueba inicial. Debe recurrirse a la investigación formativa o el análisis situacional para valorar los factores subyacentes al problema a que hace frente la intervención (por ejemplo, un mal rendimiento escolar) y fundamentar así su diseño. Por ejemplo, no tiene sentido limitarse a mejorar la asistencia escolar si el absentismo de los docentes está muy extendido.

La evaluación formativa, que puede adoptar entre otras la forma de un estudio piloto o de una prueba de concepto, examina si es viable ejecutar el programa con la suficiente aceptación y mejora la calidad de la implementación. Los resultados de un estudio piloto pueden servir para identificar modificaciones que es preciso aplicar en el diseño de un posible estudio más amplio posterior al piloto. Aunque los ensayos controlados aleatorios se han empleado en ocasiones para obtener una prueba de concepto, no deben tener ese fin en una fase temprana. Los ensayos controlados aleatorios resultan costosos; utilizarlos para evaluar un programa que todavía no se ha desarrollado plenamente supone malgastar unos recursos escasos y puede propiciar constataciones engañosas.

Los ensayos controlados aleatorios deben ser acordes con la naturaleza del programa evaluado

El uso de los ensayos controlados aleatorios resulta óptimo cuando los programas evaluados persiguen impactos claros y medibles que pueden atribuirse a una intervención o un conjunto de intervenciones concretos, y se prestan al [análisis de la vía causal](#). No son adecuados para los programas emergentes o que persiguen resultados difíciles de medir.

Se han identificado cuatro condiciones en las que la distribución aleatoria no es recomendable o viable², a saber: cuando se necesitan respuestas rápidas; cuando no se requiere mucha precisión y la pregunta causal es el objetivo más importante; cuando no es posible intervenir en la distribución, por ejemplo cuando la pregunta causal que hay que responder implica una exposición a una situación adversa; y cuando no se ha llevado a cabo un trabajo empírico previo suficiente y la intervención o el programa se encuentran en una etapa prematura.

Entre los programas que quizá no se presten a la aleatorización están los que cuentan con un número limitado de unidades de tratamiento, tales como los programas de apoyo institucional a un único organismo; o aquellos cuyas actividades y resultados previstos no se definen de manera anticipada con claridad. En ocasiones, sin embargo, un programa que a primera vista parece inadecuado para la aleatorización puede volverse propicio con un poco de imaginación. Por ejemplo, es posible evaluar un programa nacional mediante un diseño basado en el estímulo. En el marco de un programa de reforma institucional, las políticas sobre la retribución de los trabajadores podrían desarrollarse a través de un ensayo controlado aleatorio sobre una serie de paquetes de incentivos.

En aquellas situaciones en que no resulta adecuado llevar a cabo un ensayo controlado aleatorio, será preciso informar al respecto a los responsables de las decisiones pertinentes. En tales situaciones, puede recurrirse para la evaluación de impacto a un [diseño cuasiexperimental](#) (como un [emparejamiento por puntuación de la propensión](#)) o un [diseño no experimental](#) riguroso (por ejemplo, un [seguimiento de procesos](#)). Véase la Síntesis n.º 8 (Diseños y métodos cuasiexperimentales) y la Síntesis n.º 6 (Sinopsis: Estrategias de atribución causal).

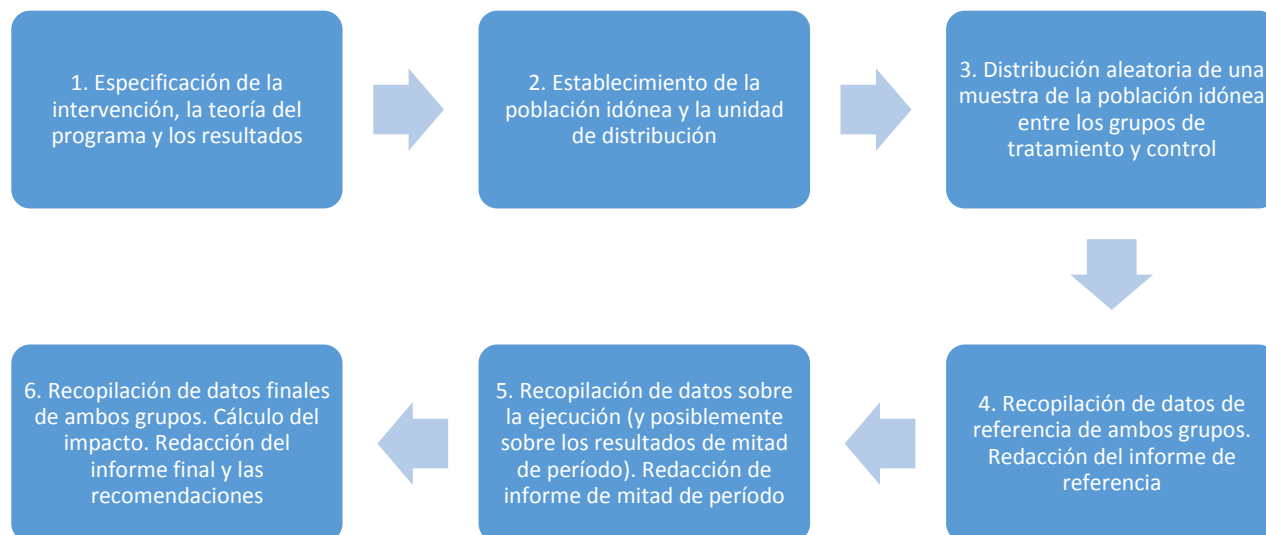
Si es demasiado pronto para llevar a cabo una evaluación de impacto, quizá sea posible y útil efectuar una evaluación que haga hincapié en los productos o resultados de los niveles inferiores de la cadena causal. La evaluación de resultados recopila información sobre los resultados (a corto plazo) que se relacionan estrechamente con los impactos de interés. Si se conoce a fondo la cadena causal y se define correctamente, esta alternativa puede resultar adecuada.

En una campaña de vacunación, por ejemplo, quizá sea preferible que la evaluación informe sobre la situación de inmunización, en lugar de esperar a recoger datos posteriores de morbilidad o mortalidad. También cabe acometer una evaluación de procesos que mejore la ejecución y aumente la utilidad de una evaluación de impacto posterior.

² Shadish, William R., et al., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston, 2002.

3. ¿CÓMO LLEVAR A CABO UN ENSAYO CONTROLADO ALEATORIO?

Gráfico 1. Resumen del proceso de los ensayos controlados aleatorios



1. Especificación de la intervención, la teoría del programa y los resultados

Como toda evaluación de impacto, el ensayo controlado aleatorio debe partir de una especificación clara de qué se va a evaluar y por qué, así como de los resultados e impactos de interés. En este sentido, resulta tremendamente útil una [teoría del cambio](#) del programa que describa lo que el programa trata de lograr y cómo —cuáles serán los procesos de cambio y cómo los propiciarán las actividades programáticas— (véase la Síntesis n.º 2 (La teoría del cambio)). El análisis de la teoría del cambio ayuda a identificar las preguntas de evaluación referentes a la cadena causal, así como a determinar qué impactos deben evaluarse. También resulta útil en esta fase aclarar cómo se emplearán en el ensayo controlado aleatorio las pruebas que se generen.

2. Determinación de la población pertinente y la unidad de distribución

Como ya se ha señalado en esta síntesis, al diseñar un ensayo controlado aleatorio es importante que la población idónea y la unidad de distribución con fines de aleatorización se determinen de manera inequívoca y se garantice la coherencia. Una de las decisiones importantes que habrá que tomar corresponde a la unidad de distribución —es decir, si en los grupos de tratamiento y control deben distribuirse aleatoriamente individuos o grupos de individuos, tales como escuelas o pueblos. Asimismo, los evaluadores deben decidir al principio del estudio qué subgrupos podrían ser de interés, de manera que el estudio tenga potencia estadística suficiente para efectuar análisis de interés sobre tales grupos.

3. Distribución aleatoria de una muestra de la población idónea entre los grupos de tratamiento y control

Los ensayos controlados aleatorios pueden diseñarse de modos diversos. Existen varias alternativas para ejecutar un ensayo controlado aleatorio sobre un programa, y el diseño del ensayo podrá determinarse en función de las características de aquel. A continuación se describen tres diseños habituales.

La aleatorización gradual implica que, con el tiempo, todas las unidades de distribución recibirán el programa si este se considera eficaz. Así pues, lo que se distribuye aleatoriamente en este caso es el momento de entrada en el programa. Los organismos de ejecución a menudo implantan los programas por etapas, y por tanto es posible elegir aleatoriamente el orden en que los participantes lo recibirán. Por ejemplo, si las limitaciones presupuestarias y logísticas impiden implantar de inmediato un programa a escala nacional, quizá sea posible seleccionar aleatoriamente las unidades que sí lo recibirán en la primera fase. Un ejemplo muy conocido de este enfoque es el programa de transferencia monetaria condicionada Progres/Oportunidades, que se llevó a cabo en México³. En su fase inicial, el programa se puso a prueba en 506 comunidades, la mitad de las cuales recibió el programa en un primer momento mientras que la otra mitad actuó como grupo de control durante dos años. Así pues, las comunidades se distribuyeron aleatoriamente en ambos grupos y disfrutaron del programa, o bien en el primer año, o bien en el tercero (es decir, las que recibieron el programa en el tercer año llevaban dos años ejerciendo de grupo de control).

La aleatorización con umbral (aumentado) puede emplearse cuando la población idónea supera la capacidad de servicio asociada a los recursos presupuestarios disponibles. Puesto que el programa no se ofrecerá a todas las personas que cumplen los criterios, seleccionar aleatoriamente a sus beneficiarios es en ocasiones la manera más justa y transparente de decidir quién participa. Estos casos son frecuentes cuando un umbral determina los criterios de selección, como por ejemplo el umbral de pobreza. Si el presupuesto disponible no permite cubrir a toda la población que reúne los requisitos, elevar ligeramente el umbral hace posible la aleatorización. Por ejemplo, si el criterio para optar a un programa de nutrición se establece en los hogares con niños de hasta 24 meses, puede incrementarse dicho umbral hasta los 30 meses. Cabe aplicar un enfoque análogo con los criterios geográficos. Si un programa prevé trabajar en 50 comunidades, pueden identificarse 100 comunidades en un primer momento y después elegir aleatoriamente 50 que disfrutarán del programa. En este caso, la técnica de la aleatorización de emparejamiento —es decir, dividir a las comunidades por parejas y asignar aleatoriamente una comunidad de cada pareja al grupo de tratamiento— dotaría de mayor solidez al diseño.

Los diseños basados en el estímulo se usan en programas y políticas universalmente disponibles, pero que no todos adoptan. Se alienta al grupo de tratamiento para que disfrute de la intervención, aunque tal estímulo no debe afectar en sí mismo a la intervención. Como ejemplo de estímulo adecuado cabe mencionar el despliegue de campañas informativas sobre un programa en curso en un conjunto determinado de poblaciones. Las poblaciones donde se llevarán a cabo esas campañas se eligen aleatoriamente entre todas aquellas donde se ha implantado el programa. A continuación, los investigadores miden el impacto del programa en los resultados de interés, comparando los resultados de las poblaciones de control y tratamiento (es este caso, aquellas que han estado expuestas a la campaña de información). Este enfoque permite calcular el impacto gracias a los diferentes índices de utilización de las poblaciones donde se llevan a cabo las campañas informativas y de aquellas donde no.

La distribución aleatoria de grupos de población entre los grupos de tratamiento y control puede efectuarse de varias maneras, entre otras:

- **Aleatorización simple:** Se elabora un listado de los individuos o emplazamientos y después se asignan a los grupos de tratamiento y control mediante números adjudicados al azar, por ejemplo mediante un generador de números aleatorios.

³ Véase Secretaría de Desarrollo Social (SEDESOL), Oportunidades, <http://www.imss.gob.mx/imss-prospera>.

- **Aleatorización por emparejamiento:** Los individuos o conglomerados se dividen en parejas a partir de una serie de características apreciables semejantes. Una de las unidades del par se asigna aleatoriamente al grupo de tratamiento; la otra, al grupo de control. El emparejamiento inicial favorece el equilibrio y reduce el tamaño de muestra necesario.
- **Distribución aleatoria estratificada:** Cuando es probable que las variables principales influyan en los resultados (por ejemplo, el nivel de ingresos o educativo), los participantes se dividen en grupos (estratos) —tales como ingresos bajos, medios y altos— y a continuación se lleva a cabo la aleatorización de cada uno de ellos. De este modo, se asegura una distribución equivalente de las variables principales entre los grupos de tratamiento y control.

Es preciso respetar el proceso de distribución aleatoria, y vigilar periódicamente su cumplimiento, a lo largo de la evaluación (véase a continuación).

4. Recopilación de datos de referencia de ambos grupos

Antes o después de la distribución aleatoria de los participantes, los evaluadores suelen realizar una encuesta inicial a partir de la cual se generan los datos en que se basarán las comparaciones finales (y quizá otras posteriores). Los datos de referencia también sirven para calibrar la equivalencia de las características iniciales de los grupos de tratamiento y control. Esa valoración de la equivalencia, también denominada comprobación de equilibrio, permite cerciorarse de que los promedios del grupo de tratamiento y el grupo de control son parecidos en diversas variables observables. Se hace principalmente para confirmar que la aleatorización se llevó a cabo correctamente. Cuando se aprecian (o se prevén) diferencias importantes, pueda estar justificado el uso de la distribución aleatoria estratificada.

Como se apuntó anteriormente, el tamaño de la muestra de la encuesta viene determinado por el cálculo de la potencia estadística. Normalmente, la encuesta inicial tiene lugar entre individuos u hogares. En ella se recogen datos sobre las características de los hogares, su situación socioeconómica, nivel educativo, salud y cualquier otro rasgo que pueda relacionarse con el programa objeto de evaluación y los impactos que persigue.

5. Recopilación de datos sobre la ejecución (y posiblemente sobre los resultados de mitad de período)

Han de recogerse datos que faciliten información sobre la ejecución, posiblemente a través de una encuesta de mitad de período que suele centrarse en los aspectos procedimentales. Esa encuesta puede servir también para obtener las estimaciones iniciales del impacto del programa, si no resulta prematuro.

Es importante cerciorarse de que las personas del grupo de control no padecen problemas de «contaminación», ya sea por la ejecución de una intervención semejante en las zonas de control, o a través de la autocontaminación, consistente en que los participantes en el estudio pasen de un grupo a otro y, de ese modo, contaminen el proceso de aleatorización inicial. También es conveniente comprobar el desgaste en los grupos, que puede propiciar resultados engañosos. Por ejemplo, si el número de personas que facilitan datos sobre los resultados es menor en el grupo de control que en el de tratamiento, se crearía un sesgo en los resultados (pues los participantes que abandonan quedan excluidos del análisis).

6. Recopilación de datos sobre los impactos

Tras la ejecución del programa se lleva a cabo una encuesta de final de período. La fecha de final de período depende de la teoría del cambio, en concreto del tiempo necesario para que se produzcan los impactos previstos. Cuando los programas prosiguen durante un período más amplio, los datos de final de

período pueden recopilarse una vez transcurrido un tiempo razonable, cuando empiecen a apreciarse los cambios en los resultados debidos a la intervención.

Por ejemplo, en una intervención nutricional que incluya el enriquecimiento con hierro, los investigadores deben esperar el tiempo suficiente para que el programa se ponga en marcha y los participantes tengan una exposición adecuada al enriquecimiento con hierro, de manera que este sea absorbido por el organismo y la incidencia de la anemia comience a disminuir (gracias al mayor consumo de hierro). Los datos de final de período se emplean para obtener estimaciones de impacto.

Si la aleatorización no tuvo demasiado éxito (según las conclusiones de las comprobaciones de equilibrio efectuadas anteriormente), habrá que emplear el método de la [diferencia en diferencias](#). Tal como se indica en la Síntesis n.º 10 (Sinopsis: métodos de recolección y análisis de datos en la evaluación de impacto), ese método mide la diferencia en el cambio que se aprecia en los resultados de los grupos de tratamiento y control. También permite valorar los efectos heterogéneos y diferenciales en los subgrupos.

El efecto del programa puede diferir entre grupos distintos, tales como hombres y mujeres, ricos y pobres, o personas con formación o sin ella. Comparar los resultados de esas categorías, a partir de los grupos de control y tratamiento, puede ayudar a calcular los impactos en dichos subgrupos. Para ello, se compara la diferencia en diferencias en los efectos de varios subgrupos de la población idónea; por ejemplo, entre los hombres, las mujeres y los niños.

En función de la naturaleza del programa, quizá los evaluadores también estén interesados en calcular los efectos a largo plazo de la intervención mediante una encuesta posterior al final de período. El momento en que se lleven a cabo tales encuestas de seguimiento dependerá de la celeridad con que se prevé que la intervención obtenga resultados. Por ejemplo, los programas de transferencia de alimentos suelen generar beneficios nutricionales con relativa rapidez, mientras que las intervenciones dirigidas a modificar las actitudes, normas y conductas vigentes normalmente tardan más tiempo en lograr resultados.

4. CUESTIONES ÉTICAS Y LIMITACIONES PRÁCTICAS

Los ensayos controlados aleatorios plantan una serie de preocupaciones éticas y prácticas semejantes a las de todas las evaluaciones (véase la Síntesis n.º 1 (Sinopsis de la evaluación de impacto), pero esta sección hace hincapié exclusivamente en las cuestiones específicas de los ensayos controlados aleatorios.

Estos hacen frente a cuestiones éticas propias relacionadas con su naturaleza experimental, por las que es importante consultar a los participantes en la prueba y conocer y hacer frente a sus anhelos, así como equilibrar los riesgos y beneficios conexos. Las preocupaciones éticas en torno a la experimentación son todavía más prominentes en el caso de los ensayos controlados aleatorios en los que el grupo de control no recibe ninguna intervención. La situación de desventaja que podría generarse hace muy importante que el proceso de aleatorización sea transparente, especialmente cuando tiene lugar entre individuos. El evaluador es responsable de velar por que no surjan tensiones entre los grupos de tratamiento y control.

Una de las formas de mitigar esta posibilidad consiste en explicar con claridad el propósito de la aleatorización. El diseño gradual puede aliviar estas preocupaciones éticas, si es posible garantizar que la intervención llegará al grupo de control en caso de resultar eficaz.

Otra preocupación ética se deriva, en un primer momento, de la propia necesidad de llevar a cabo un ensayo controlado aleatorio. Cuando no existen dudas razonables sobre los beneficios y la rentabilidad de un programa, no es preciso efectuar una evaluación exhaustiva (de ningún tipo). Efectuar un seguimiento del impacto quizá sea un método más adecuado para evaluar si el programa sigue brindando los resultados previstos a lo largo del tiempo. No obstante, si surgen interrogantes sobre la eficacia de un programa y solo se dispone de recursos limitados para su implantación, puede considerarse más ético distribuir aleatoriamente a los participantes en el programa, con el propósito de ampliarlo a toda la población si resulta eficaz.

También es importante encarar con prudencia la recopilación de datos sobre el grupo de control. Los evaluadores deben preocuparse por dar un uso adecuado al tiempo de las personas no beneficiarias. Es lógico compensar a los encuestados por su tiempo, aunque ha de hacerse de manera que no afecte a los resultados (por ejemplo, no debe animarse a los individuos a responder de una forma determinada).

5. ¿QUÉ OTROS MÉTODOS FUNCIONAN BIEN CON ESTE?

El ensayo controlado aleatorio es un diseño de investigación que persigue expresamente dar respuesta a una serie de preguntas sobre causalidad y atribución. Su diseño debe enmarcarse en un plan general de evaluación que comprenda también otros métodos para la recopilación de datos (tales como entrevistas, observaciones y mediciones directas) y su análisis.

Es recomendable llevar a cabo una investigación o evaluación formativa antes del ensayo controlado aleatorio, a fin de valorar la viabilidad de la implantación de un programa y de mejorar la ejecución.

Asimismo, es conveniente que el ensayo controlado aleatorio, o bien conste de un componente que examine los procesos de implantación, o bien se complemente con una evaluación de procesos. Mientras que el ensayo controlado aleatorio se plantea la cuestión contrafáctica, la evaluación de procesos se pregunta cómo se ejecutó el programa, normalmente tomando como referencia datos tanto cuantitativos como cualitativos. Así pues, la evaluación de procesos responde a cuestiones de toda la cadena causal y resulta de ayuda para explicar los motivos del impacto del programa. Los datos recopilados complementarán, y podrán emplearse para verificar (o desmentir), los datos del propio sistema de seguimiento del programa.

6. PRESENTACIÓN DE LOS RESULTADOS Y ANÁLISIS

Es preciso llevar a cabo una garantía de calidad de los ensayos controlados aleatorios que asegure la calidad del estudio, e importante redactar con el suficiente grado de detalle la metodología y las constataciones. Igualmente fundamental es no centrarse únicamente en la metodología al describir el ensayo controlado aleatorio; también hay que describir la intervención objeto de evaluación. Esa información puede obtenerse en la evaluación de procesos correspondiente. Describir pormenorizadamente la intervención permite vincular la teoría del cambio con el análisis de las constataciones. Asimismo, debe describirse en detalle la teoría del cambio al comunicar las constataciones del ensayo controlado aleatorio.

Cuando se desarrolla la metodología, es recomendable explicar tanto la muestra como el método de distribución aleatoria. En esa explicación, es importante indicar el número de conglomerados y el número de hogares o individuos que forman parte de los grupos de tratamiento y control. El informe debe constar asimismo de tablas sobre las comprobaciones de equilibrio (descritas anteriormente).

Las estimaciones de impacto pueden presentarse por medio de un análisis de la diferencia en diferencias, cuyas constataciones cabe exponer en relación con toda la muestra y también con los subgrupos, con vistas a analizar posibles efectos heterogéneos.

Las constataciones se relacionan posteriormente con la teoría del cambio. ¿Coincide el análisis con la teoría del cambio? Si no es así, ¿qué supuesto subyacente a la teoría del cambio no se cumplió? Han de explicarse otras constataciones del estudio. ¿Qué motivos, tanto propios como ajenos a la teoría del cambio, pueden haber propiciado los resultados? Este análisis sirve en ocasiones a los evaluadores para identificar constataciones concretas pertinentes en materia de políticas, que deben exponerse en la conclusión del informe y vincularse expresamente con el análisis de datos. En la mayoría de los casos, también es importante comentar si los datos son extrapolables a otros contextos y, de ser así, a cuáles.

7. EJEMPLO DE BUENAS PRÁCTICAS

El ensayo de ampliación del desarrollo en la primera infancia en el Pakistán⁴ financiado por UNICEF se proponía evaluar la eficacia y viabilidad de la integración de las intervenciones en favor de la primera infancia con miras a fortalecer y mejorar los resultados de salud. El ensayo se sirvió del programa en curso de trabajadoras sanitarias para facilitar los componentes de la intervención. El estudio se basó en un diseño factorial aleatorio por conglomerados y tuvo lugar en las zonas rurales del Pakistán.

Los conglomerados (definidos como zonas de captación de las trabajadoras sanitarias) se asignaron aleatoriamente a los grupos siguientes: el **grupo de control**, que recibió servicios básicos de salud y nutrición del programa de trabajadoras sanitarias; el **grupo con una nutrición mejorada**, que recibió asesoramiento nutricional, consejos sensibles sobre alimentación y un suplemento nutricional (Sprinkles®) para los niños de 6 a 24 meses; el **grupo de desarrollo en la primera infancia**, al que se brindó asesoramiento sobre estimulación y atención para el desarrollo, en particular mediante prácticas tuteladas integradas, visitas a domicilio mensuales y sesiones de grupo; y, por último, el **tercer grupo de tratamiento**, que recibió una combinación de las intervenciones en favor del desarrollo en la primera infancia y de mejora de la nutrición. Las trabajadoras sanitarias facilitaron las intervenciones a todas las familias de su zona de captación en las que había un niño menor de 24 meses.

Se midieron una serie de resultados en favor de los niños y se registraron sus datos y los de sus familias en varios momentos, hasta que el niño cumplió los 24 meses. El equipo de recopilación de datos trabajó al margen del equipo de apoyo a la intervención y sin saber en qué consistía esta. Así se minimizaron los posibles sesgos en la evaluación. El estudio concluyó que, a los 12 meses de edad, los niños de los 3 grupos de intervención obtuvieron puntuaciones en los planos cognitivo, lingüístico, locomotor y socioemocional significativamente superiores a las del grupo de control. Los grupos en los que se integró el desarrollo en la primera infancia y la mejora de la nutrición presentaron puntuaciones en los ámbitos cognitivo y lingüístico significativamente superiores a las del grupo que solo recibió nutrición mejorada. A los 24 meses de edad, los 3 grupos de intervención obtuvieron puntuaciones en los planos cognitivo, lingüístico y locomotor significativamente superiores a las del grupo de control, pero los dos grupos expuestos al desarrollo en la primera infancia tuvieron mejores resultados que los que solo disfrutaron de la nutrición mejorada.

Tenemos aquí un ejemplo de buenas prácticas en el empleo del ensayo controlado aleatorio, tanto en términos de adecuación a la situación como en su modo de ejecución y uso.

El programa se definió correctamente y estableció objetivos claros. Se elaboró un listado de comunidades aptas que se distribuyeron aleatoriamente entre los grupos de tratamiento y control. La recopilación y el análisis de datos se fundamentaron en una teoría del cambio sólida, y gracias a ello los evaluadores pudieron valorar las vías por las que la intervención había logrado sus objetivos.

8. EJEMPLOS DE POSIBLES PROBLEMAS

Mantener la integridad del diseño: Incluso cuando se aplica una distribución aleatoria, pueden surgir varios problemas: 1) utilización escasa de la intervención; 2) incumplimiento de los procedimientos previstos; 3) contaminación del grupo de control a causa de otras intervenciones con incidencia en resultados semejantes o a través de la autocontaminación; y 4) cambios en el diseño o el lugar donde se ejecuta el programa objeto de evaluación. La mayoría de esos problemas pueden solventarse en la fase de análisis, pero es preciso que los evaluadores recopilen los datos necesarios para ser conscientes de ellos y hacerles frente.

⁴ Yousafzai, A. K., *et al.*, «The Pakistan Early Child Development Scale Up (PEDS) Trial: Outcomes on child development, growth and health», informe sobre los datos relativos a los resultados del ensayo, oficina de UNICEF en el Pakistán, 2012.

La utilización escasa afecta a los programas que carecen de interés para sus beneficiarios previstos o que estos no comprenden adecuadamente. Por ejemplo, es frecuente que los planes de seguros tengan una aceptación escasa. Para conocer los motivos de un índice de utilización bajo suelen requerirse datos cualitativos.

Encontramos un ejemplo de incumplimiento en un ensayo controlado aleatorio efectuado en China⁵, en el que se facilitaron gafas a un grupo de estudiantes de secundaria. Se observó que su utilización también había aumentado en el grupo de control. Las indagaciones posteriores descubrieron que los médicos que llevaban a cabo las pruebas oculares habían entregado las gafas que el grupo de tratamiento no había utilizado a otros estudiantes de la zona de control, en lo que constituye un ejemplo de autocontaminación. El estudio se basó en un diseño de emparejamiento, de manera que fue posible descartar los pares en los que el control se había contaminado.

No corregir desviaciones estándar al trabajar con un diseño por conglomerados: se trata de un error técnico frecuente, que incrementa de manera artificial la potencia estadística, de manera que a veces se concluye erróneamente que un programa funciona, cuando sucede todo lo contrario. Por ejemplo, imaginemos que se distribuyen aleatoriamente una serie de escuelas entre varios grupos de tratamiento, pero al analizar los resultados —al comparar los resultados de aprendizaje a partir de las calificaciones de las pruebas efectuadas por los niños de los grupos de tratamiento y control— no se controla la segmentación de las escuelas, con lo que se obtiene una estimación excesiva del impacto. Todos los programas informáticos de estadística que se emplean en las evaluaciones de impacto permiten corregir este aspecto, así que los investigadores no tienen excusa.

Atención excesiva al efecto medio del tratamiento: el ensayo controlado aleatorio facilita una estimación objetiva del efecto medio de un programa. No obstante, se trata de la constatación de menor interés para los responsables de las políticas, que suelen mostrarse especialmente interesados en el grado de eficacia de un programa con subgrupos concretos, sobre todo cuando los programas hacen frente a problemas de equidad. Por ejemplo, una evaluación del programa Early Head Start⁶, dedicado a la intervención en la primera infancia en los Estados Unidos, concluyó que este, aunque eficaz en término medio, resultaba nocivo para las familias más vulnerables. Un informe sencillo que hiciera hincapié en su efecto medio proporcionaría una orientación engañosa a los responsables de las políticas y los proveedores de servicios. Una de las formas de responder a este problema consiste en informar sobre los resultados de distintos subgrupos en los que la intervención puede tener un impacto diferente. Hay que señalar que este tipo de diseño debe determinarse en la fase de diseño del estudio, y los cálculos de potencia estadística han de efectuarse de manera acorde.

Oposición a la distribución aleatoria: sucede con frecuencia que el personal del organismo de ejecución se opone a la distribución aleatoria. Contar con la aprobación de la dirección quizá no baste para disfrutar de la cooperación de los trabajadores sobre el terreno. El estudio sobre las gafas en China, en el que los médicos entregaron los anteojos al grupo de control, constituye un ejemplo de ello.

⁵ Glewwe, Paul, *et al.*, «The Impact of Eyeglasses on the Academic Performance of Primary School Students: Evidence from a Randomized Trial in Rural China», ponencia, Universidad de Minnesota y Universidad de Michigan, 2006. Véase <http://ageconsearch.umn.edu/bitstream/6644/2/cp06gl01.pdf>.

⁶ Westthorp, Gill, «Using complexity-consistent theory for evaluating complex systems», *Evaluation*, 18 (4), 2012, págs. 405-420.

9. LECTURAS Y ENLACES CLAVE

Ambroz, Angela y Marc Shotland, «Randomized Control Trial (RCT)», página web, BetterEvaluation, 2013. Véase <http://betterevaluation.org/plan/approach/rct>.

Bloom, Howard, «The Core Analytics of Randomized Experiments for Social Research», MDRC Working Papers on Research Methodology, MDRC, Nueva York, 2006. Acceso libre: http://www.mdrc.org/sites/default/files/full_533.pdf.

Duflo, Esther, *et al.*, «Using Randomization in Development Economics Research: A Toolkit», Departamento de Economía del Instituto Tecnológico de Massachusetts y Abdul Latif Jameel Poverty Action Lab, Cambridge, 2006. Véase <http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf>.

Gertler, Paul J., *et al.*, *Impact Evaluation in Practice*, Banco Mundial, Washington D. C., 2010. Véase http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf.

Glennerster, Rachel y Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*, Princeton University Press, Princeton, 2013.

Raudenbush, S. W., *et al.*, *Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01)*, software, 2011. Disponible en www.wtgrantfoundation.org.

Spybrook, J., *et al.*, *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0*, documentación de software, 2011. Disponible en www.wtgrantfoundation.org.

Urbaniak, Geoffrey C. y Scott Plous, *Research Randomizer (Version 4.0)*, recursos para la distribución aleatoria y el muestreo aleatorio, software, 2013. Disponible en <http://www.randomizer.org/>.

White, Howard, «An introduction to the use of randomised control trials to evaluate development interventions», *Journal of Development Effectiveness*, 5 (1), 2013, págs. 30-49. Acceso libre: <http://www.tandfonline.com/doi/pdf/10.1080/19439342.2013.764652>.

GLOSARIO

<u>Análisis de la vía causal</u>	<i>Análisis basado en una vía causal (también conocida como marco analítico) que describe relaciones directas e indirectas entre las variables independientes (las intervenciones) y dependientes (productos, resultados e impactos).</i>
<u>Contaminación</u>	<i>Inclusión de un individuo o grupo de entrevistados en un grupo de tratamiento (intervención) que no representa a la población o que en teoría no debe recibir el tratamiento. Puede suceder cuando los participantes o sujetos del grupo de control reciben el tratamiento de forma involuntaria, de manera que se reducen los efectos del tratamiento en las mediciones de resultados.</i>
<u>Desgaste</u>	<i>Situación en la que los participantes en el estudio abandonan la muestra entre dos rondas de recopilación de datos. El desgaste puede poner en peligro la validez interna de un estudio y alterar la composición de la muestra.</i>
<u>Diferencia en diferencias</u>	<i>El método de la diferencia en diferencias, también conocido como de «diferencia doble», compara los cambios en los resultados en el curso del tiempo entre los grupos de tratamiento y de comparación a fin de estimar el impacto.</i>
<u>Diseño basado en el estímulo</u>	<i>Diseño experimental que no asigna aleatoriamente a los participantes a una intervención propiamente dicha, sino que les hace llegar material promocional o información adicional sobre los beneficios de una intervención a su alcance, a fin de fomentar su participación. Pueden emplearse diseños basados en el estímulo cuando un programa universalmente disponible no ha sido adoptado por todos. Véase: diseño no experimental.</i>
<u>Diseño cuasiexperimental</u>	<i>Diseño de investigación o evaluación en el que los participantes no se distribuyen de forma aleatoria entre las condiciones de tratamiento, sino que los grupos de comparación se crean por medios estadísticos. Difiere del experimento controlado (clásico) en que el tratamiento o intervención no se distribuye de manera aleatoria.</i>
<u>Diseño del ensayo controlado aleatorio por conglomerados</u>	<i>Diseño experimental en el que la unidad de distribución consta de múltiples unidades de tratamiento, en lugar de un único sujeto. Por ejemplo, las intervenciones educativas suelen distribuirse por escuelas, aunque la intervención tiene lugar entre los profesores, en las clases o con cada niño, y sus efectos se miden en los niños. Véase: ensayos controlados aleatorios.</i>
<u>Diseño experimental</u>	<i>Diseño de investigación o evaluación con dos o más grupos seleccionados de forma aleatoria (un grupo experimental y un grupo de control) en los que el investigador controla o introduce una intervención (por ejemplo, un nuevo programa o política) y mide su impacto en la variable dependiente al menos dos veces (medición anterior y posterior al ensayo). En concreto, los ensayos controlados aleatorios —que tienen su origen en el contexto clínico y se conocen como la «regla de oro» de la investigación médica y sanitaria— se suelen utilizar para responder a las preguntas de investigación de la evaluación, que tratan de evaluar la eficacia de las intervenciones de un programa o política en entornos de desarrollo.</i>

<u>Diseño no experimental</u>	<i>Tipo de diseño de investigación que no incluye un grupo de control o testigo ni una evaluación de referencia. Por consiguiente, varios factores impiden que un efecto observado se atribuya a la intervención. Véase: diseño de investigación experimental, diseño de investigación cuasiexperimental.</i>
<u>Emparejamiento por puntuación de la propensión</u>	<i>Método cuasiexperimental que coteja los individuos u hogares en tratamiento con otros semejantes a efectos comparativos, y posteriormente calcula la diferencia media en los indicadores de interés.</i>
<u>Encuesta de final de período</u>	<i>Encuesta que se lleva a cabo cuando concluye la intervención, por lo general con el propósito de comparar los resultados con la encuesta inicial. Término conexo: encuesta inicial.</i>
<u>Grupo de tratamiento</u>	<i>Los sujetos o participantes expuestos a la variable independiente; también llamado «grupo experimental» o «grupo de intervención».</i>
<u>Seguimiento de procesos</u>	<i>Enfoque de la inferencia causal basado en los casos que consiste en desarrollar hipótesis alternativas y posteriormente recabar pruebas (pistas) en un caso para determinar si son compatibles o no con las hipótesis.</i>
<u>Teoría del cambio</u>	<i>Explica cómo se entiende que las actividades produzcan una serie de resultados que contribuyen a lograr los impactos finales previstos. Puede elaborarse para cualquier nivel de intervención: un acontecimiento, un proyecto, un programa, una política, una estrategia o una organización.</i>