Innovamat

¿Funciona o no funciona? Cómo entender la evidencia de un programa de matemáticas

Una guía para leer la evidencia sobre la eficacia de un programa de matemáticas

Marc Colomer
Director de Investigación del Innovamat Research Lab

Carles Aulés
Especialista de Investigación del Innovamat Research Lab

Caterina Calsamiglia
ICREA Research Professor y responsable de la unidad de Computational Social Science and Humanities del Barcelona Supercomputing Center (BSC)

Innovamat Research Series

Resumen

Cuando una escuela incorpora un nuevo programa de matemáticas, es común preguntarse: ¿Funciona? Este capítulo recorre las piezas que hacen falta para evaluar un programa educativo y leer un estudio de impacto. Se describen tres ideas principales:

- Demostrar si un programa educativo funciona en abstracto no es posible. La pregunta que sí podemos hacer es: «¿Qué diferencia supone utilizar este programa en lugar de la alternativa disponible, para un resultado concreto, una determinada población y un contexto específico?»

- No toda evidencia tiene el mismo peso y es igual de informativa. La rigurosidad de un estudio de impacto depende de la capacidad que tiene de estimar qué les habría pasado a esos mismos alumnos sin el programa, y eso exige construir una comparación justa con alumnos equivalentes que no lo recibieron.

- Un buen estudio de evaluación de impacto requiere también una buena teoría del cambio: ¿qué debería cambiar, a través de qué mecanismo y bajo qué condiciones? Es importante definir qué se espera que pase cuando se usa el programa, qué debería cambiar y entender la implementación como parte fundamental del proceso.

El capítulo se cierra con ocho preguntas para leer cualquier estudio y con las dudas más frecuentes de quien recibe un informe de eficacia.

Cuando una escuela incorpora un nuevo programa de matemáticas[1], aparece una pregunta recurrente entre docentes y familias: «De acuerdo, pero... ¿funciona?».

La pregunta es del todo legítima y cada vez parece más necesario medir si aquello que hacemos en educación tiene impacto, y si ese impacto es positivo o negativo. Así como en medicina nadie se plantea aplicar un tratamiento sin que antes se hayan investigado sus efectos, en educación esta idea aún no ha calado del todo. Afortunadamente, se extiende cada vez más la costumbre de evaluar la eficacia de las prácticas y los recursos que entran en la escuela, y de decidir a partir de evidencias.

Eso nos lleva a un problema que no es menor. En algunos casos, la palabra «evidencia» se ha convertido en un argumento de venta. Los materiales llegan a los centros acompañados de estudios, porcentajes y sellos que prometen mejoras demostradas del aprendizaje. Es una buena noticia que cada vez se pida más evidencia empírica. Conviene, eso sí, saber qué pregunta responde esa evidencia y hasta qué punto la respuesta nos sirve a nosotros.

Para empezar, no podemos responder si un programa no *funciona* en abstracto. Lo que sí podemos responder es una pregunta más concreta. **¿Qué diferencia supone utilizar este programa en lugar de la alternativa disponible, para un resultado concreto, una determinada población y un contexto específico?**

Esta forma de plantear la pregunta importa. El efecto de un programa depende de con qué lo comparemos, de qué resultado midamos, de quién participe en el estudio y de las condiciones en las que se aplique. Un resultado positivo nos dice lo que ocurrió allí, y para saber si podemos esperar algo parecido aquí hace falta bastante más.

A lo largo de este capítulo veremos cómo se construye una comparación que permita estimar el efecto de un programa, qué papel desempeñan la aleatorización y otros diseños de evaluación, qué significa medir una mejora en matemáticas, cómo interpretar el tamaño y la precisión de un efecto y qué condiciones explican que un programa funcione en unos sitios y no en otros. El objetivo es aprender a leer un resultado de impacto preguntándonos **qué diferencia estamos estimando, cómo la hemos estimado y hasta qué punto podemos esperar algo parecido en nuestro contexto.**

Al final de este capítulo

cuando alguien te diga que un programa de matemáticas «funciona», sabrás qué preguntas hacer: con qué se comparó, qué aprendieron realmente los alumnos, cuánto mejoraron y en qué condiciones se consiguió ese resultado. Y, sobre todo, podrás valorar qué te dice esa evidencia sobre lo que podría ocurrir en tu aula.

[1] Entendemos un programa de matemáticas como un paquete de recursos educativos que combina materiales concretos (libros de texto, guías para docentes, materiales manipulativos, etc.) con servicios que acompañan la implementación (formación profesional, asesoramiento, etc.).

1 | Evaluar es construir una comparación justa

Imaginemos que una escuela pone en marcha un nuevo programa de matemáticas. Al cabo de un año, sus estudiantes han mejorado mucho los resultados. ¿Podemos decir que el programa ha mejorado su rendimiento?

No. Los resultados pueden haber cambiado por muchas razones. Quizá haya llegado un nuevo docente, la cohorte de estudiantes sea diferente, se hayan dedicado más horas a matemáticas o el equipo docente haya recibido formación adicional. Observar que los resultados han mejorado nos dice que algo ha cambiado; averiguar qué lo ha causado es otra cosa.

Este problema no es exclusivo de la educación. Imaginemos que una farmacéutica anuncia su nuevo jarabe diciendo que el 95 % de las personas que lo han tomado se han curado del resfriado al cabo de una semana. ¿Basta para concluir que el jarabe funciona? Muchas personas se recuperan de un resfriado en una semana sin tomar nada. Para saber si el jarabe ha tenido algún efecto necesitamos saber qué habría ocurrido con esas mismas personas si no lo hubieran tomado.

Pongamos otro ejemplo. Imagina que un estudio concluye que quienes utilizan más crema solar son también quienes sufren más quemaduras del sol. ¿Provoca la crema las quemaduras? Lo más probable es que esas personas pasen muchas más horas al sol, lo que explicaría la relación entre los resultados.

El problema es que observar que *X* y *Y* ocurren juntas no nos indica necesariamente que *X* haya causado *Y*. En investigación, esta diferencia entre asociación y causalidad es fundamental, y se resume en una frase muy repetida: **correlación no implica causalidad**.

> Lo que intenta un buen estudio de eficacia es averiguar si hacer *X* —por ejemplo, usar un programa de matemáticas— ha causado un cambio en *Y* —por ejemplo, el aprendizaje del alumnado—. No basta con observar que las dos cosas ocurren a la vez.

Volvamos a la escuela. Supongamos que el nuevo programa se introduce al mismo tiempo que llega un profesor de matemáticas con mucha experiencia y especialmente motivado. Al terminar el curso, los resultados mejoran. ¿Ha funcionado el programa? ¿Ha sido el profesor? ¿Han contribuido ambos?

Con los datos de ese único curso es imposible saberlo. Entre el «antes» y el «después» han cambiado demasiadas cosas. La comparación que de verdad nos interesa es otra. ¿Cómo les habría ido a esos mismos alumnos si no hubieran recibido el nuevo programa?

Ese escenario alternativo es lo que llamamos el **contrafactual** (Gertler et al., 2016), y tiene una propiedad incómoda. Nunca podemos observarlo directamente, porque un mismo alumno no puede recibir el programa y, al mismo tiempo, no recibirlo.

> **Contrafactual.** Qué les habría pasado a estos mismos alumnos si no hubieran recibido el programa o intervención. No se puede observar nunca, y toda la credibilidad de un estudio depende de lo bien que consiga estimarlo.

¿Cómo podemos aproximar algo que no ha ocurrido?

Con una comparación. Buscamos un grupo de alumnos lo más parecido posible al que recibió el programa, pero que no lo recibió. Lo que les ocurra nos sirve como referencia para aproximar qué habría pasado con el primer grupo en ausencia del programa. Este es el **grupo de comparación o de control**[2].

[2] Utilizaremos la palabra *control* cuando hablemos de estudios experimentales, mientras que *comparación* es más adecuada para los cuasi-experimentales o correlacionales.

Volvamos al ejemplo del jarabe. Si el 95 % de quienes toman el medicamento se recupera en una semana, la cifra por sí sola nos dice muy poco. Si entre personas comparables que no lo toman se recupera el 93 %, probablemente el jarabe esté haciendo poco. Si solo se recupera el 40 %, tendríamos una señal mucho más clara de que el medicamento está teniendo un efecto. La cifra del 95 % empieza a ser informativa cuando tenemos con qué compararla.

La cuestión clave pasa a ser, entonces, cómo construimos una comparación suficientemente justa. Cuando podemos decidir por sorteo quién recibe el programa, conseguimos que las características de los dos grupos sean comparables en promedio, incluso en aspectos que no sabemos medir. Cuando el sorteo no es posible, necesitamos otros diseños que construyan una comparación igualmente creíble. Esta es la lógica que hay detrás de los experimentos aleatorizados y de los diseños cuasi-experimentales.

> «La cuestión clave pasa a ser, entonces, cómo construimos una comparación suficientemente justa.»

La comparación nos da respuestas

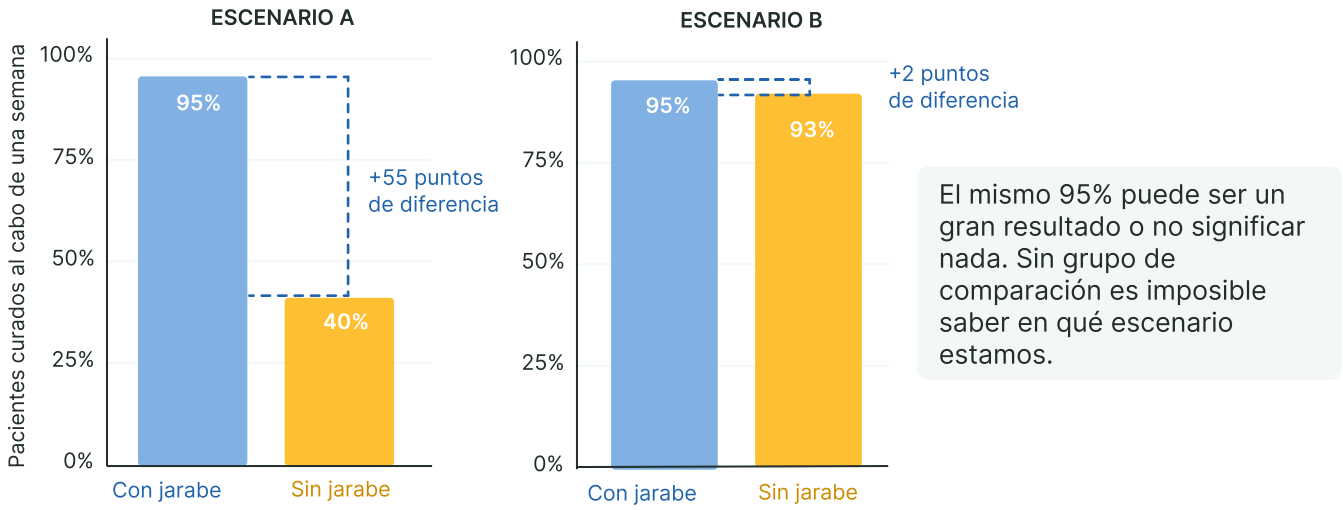

Figura 1. Simulación del estudio del jarabe con grupo de comparación y dos escenarios distintos. En el escenario A, es poco común que los pacientes se curen sin tomar el jarabe. En el escenario B, en cambio, casi todos los pacientes (un 93 %) se curan naturalmente sin tomar nada. Sin una comparación justa, no podemos interpretar el valor del 95 %.

Que las notas suban después de adoptar un nuevo programa de matemáticas todavía no demuestra que el programa funcione, porque entre el antes y el después cambian muchas más cosas que pueden ayudar a explicar la mejora. Evaluar el impacto causal es estimar qué les habría pasado a esos mismos alumnos sin el programa, y eso exige construir una comparación justa con alumnos equivalentes que no lo recibieron.

2 | ¿Cómo construimos una comparación justa?

El azar como herramienta

La forma más sencilla de construirla es utilizar el azar. Imaginemos que tenemos un grupo de escuelas interesadas en probar un nuevo programa de matemáticas. En lugar de dejar que cada escuela decida si lo utiliza, sorteamos cuáles lo recibirán y cuáles continuarán con la alternativa habitual. Conviene fijarse en dónde actúa el azar. Decide quién recibe el programa y quién sirve de comparación, entre escuelas que ya habían aceptado participar.

Si tenemos suficientes escuelas, el azar hace que, en promedio, los dos grupos sean comparables antes de empezar. Algunas escuelas tendrán docentes más experimentados y otras menos, algunas tendrán alumnado con mejores resultados iniciales y otras peores, algunas estarán más motivadas y otras menos. Como la asignación se ha hecho al azar, ninguna de estas características debería concentrarse sistemáticamente en uno de los dos grupos.

Y esto incluye características que ni siquiera sabemos medir. Podemos comparar los centros según el nivel socioeconómico de sus alumnos, sus resultados anteriores o el tamaño de la escuela, y habrá siempre otras diferencias más difíciles de observar, como las expectativas del equipo docente, la implicación de las familias o la cultura del centro. **La gran ventaja del azar es que las equilibra en promedio sin que necesitemos conocerlas.** Así, si después del programa los alumnos que lo han recibido obtienen mejores resultados que los del grupo de comparación, tenemos una razón mucho más sólida para atribuir esa diferencia al programa.

> La aleatorización hace que, en promedio, las diferencias de partida entre los dos grupos sean independientes de quién recibe el programa. Escuela por escuela seguirán siendo distintas, y aun así la diferencia posterior puede interpretarse como un efecto causal con mucha más confianza.

Este es el principio de un **experimento aleatorizado**, o *randomized controlled trial* (RCT). Es la misma lógica que se utiliza en un ensayo clínico, donde algunas personas reciben el tratamiento y otras un placebo o la alternativa habitual, y la diferencia entre ambos grupos permite estimar el efecto del tratamiento. Pero, en educación, la aplicación tiene algunas particularidades.

Los experimentos educativos tienen sus propias dificultades

Muchas veces no podemos sortear alumno por alumno. Si el programa modifica la manera de enseñar, lo habitual será asignar aulas o escuelas enteras, porque de lo contrario un mismo docente tendría que utilizar métodos diferentes con alumnos de la misma clase y lo que ocurre con unos podría afectar a los otros.

Además, a diferencia de un ensayo médico, normalmente todos saben qué programa están utilizando. El docente sabe que está enseñando con materiales nuevos y el alumnado puede saber que participa en una intervención. Parte de lo que observamos puede venir, entonces, de la novedad, de una mayor atención o del esfuerzo adicional que acompaña a cualquier experimento.

Y hay una última diferencia importante. Un material curricular es mucho más que un material. Lo que ocurre en el aula depende de cómo lo utiliza el docente, de cómo responde el alumnado y de las condiciones del centro, de manera que dos escuelas pueden recibir exactamente el mismo programa y aplicarlo de maneras muy diferentes. Al leer los resultados de un experimento conviene recordar, por tanto, qué se ha evaluado exactamente. El estudio nos dice **qué diferencia produjo ese programa, aplicado de esa manera, en esa población y en ese contexto, frente a la alternativa** utilizada como comparación.

¿Y si no podemos utilizar el azar?

No siempre podemos asignar un programa aleatoriamente. Quizá todas las escuelas quieren utilizar el programa y no se quieren esperar, quizá la administración lo haya implantado ya en determinados centros o quizá existan razones éticas o prácticas que hagan imposible un sorteo.

En estos casos los investigadores construyen la comparación de otras maneras, con los llamados **diseños cuasi-experimentales**. La lógica sigue siendo exactamente la misma, encontrar un grupo que nos permita aproximarnos lo mejor posible a lo que habría ocurrido sin el programa.

Imaginemos que una administración decide ofrecer un programa primero a las escuelas que están por debajo de un determinado resultado en una prueba. Los investigadores pueden comparar las escuelas situadas justo por encima y justo por debajo del umbral, si tienen razones para pensar que son suficientemente similares y que quedar a un lado u otro del límite es lo que determina la participación. Otros estudios aprovechan cambios de política, implantaciones progresivas o reglas administrativas que generan grupos comparables.

Un buen diseño cuasi-experimental puede proporcionar evidencia causal muy sólida. Cuando la regla que forma los grupos es convincente y los supuestos que necesita el diseño se comprueban explícitamente, la confianza que merece se acerca mucho a la de un experimento. Lo que cambia es dónde ponemos la atención al leerlo. Sin la protección que ofrece el sorteo, hay que examinar con más cuidado por qué podemos considerar comparables los dos grupos y qué supuestos necesita el diseño.

El peligro de comparar grupos que se han elegido a sí mismos

Volvamos al ejemplo de las escuelas que se ofrecen voluntariamente para probar un programa. Supongamos que al cabo de un año sus resultados han mejorado más que los de las escuelas que no participaron. ¿Ha funcionado el programa?

Quizá sí, y también puede haber ocurrido otra cosa. Las escuelas que se ofrecieron voluntariamente pueden tener docentes especialmente motivados, equipos directivos que dan mucha importancia a las matemáticas o más capacidad para introducir cambios. También podría ocurrir lo contrario, que sean centros con resultados especialmente bajos que estaban buscando activamente una solución.

Si esas características también afectan a los resultados, la diferencia que observamos puede reflejar quién decidió participar. Esto es lo que llamamos **sesgo de selección**, y es especialmente difícil de detectar porque algunas de las diferencias relevantes pueden no aparecer en los datos. Podemos hacer que los dos grupos sean muy parecidos en tamaño, titularidad, nivel socioeconómico o resultados iniciales y que difieran aun así en algo importante que no estamos midiendo.

> Comparar a quienes han recibido un programa con quienes no lo han recibido parece una comparación natural y puede ser muy engañosa. Si las personas, escuelas o docentes que eligen participar son diferentes de los que no participan, esas diferencias pueden confundirse con el efecto del programa.

La calidad de la comparación determina la fuerza de la evidencia

Por eso, cuando leemos un estudio, una de las primeras preguntas debería ser cómo se decidió quién recibía el programa y quién no. Un sorteo nos da una base muy sólida; cualquier otra regla nos obliga a entender por qué permite pensar que los grupos son comparables.

Esto explica por qué estudios que utilizan la misma palabra, «evidencia», permiten grados de confianza muy distintos. Un estudio puede observar que los alumnos que utilizan un programa obtienen mejores resultados; si esos alumnos fueron seleccionados de una manera que los hace diferentes de los demás, seguimos sin saber cuánto de esa diferencia se debe al programa. Otro estudio puede utilizar una

comparación mucho más rigurosa y encontrar una diferencia más pequeña. Paradójicamente, **el segundo resultado puede ser mucho más informativo**. Valorar la credibilidad de la comparación es, por tanto, el paso previo a mirar el tamaño del efecto.

Una forma de ordenar la evidencia

Esta diferencia entre diseños es también la razón por la que algunos sistemas clasifican la evidencia según la confianza que permite tener en una afirmación causal. Uno de los marcos más conocidos es el de la Every Student Succeeds Act (ESSA, 2015) de Estados Unidos, que distingue cuatro niveles: evidencia fuerte, moderada, prometedora y fundamentación. Los niveles superiores exigen diseños que permitan una inferencia causal más convincente. El nivel más bajo recoge programas que cuentan con una justificación basada en investigación y que todavía no han sido suficientemente evaluados.

Los cuatro peldaños de la escala ESSA

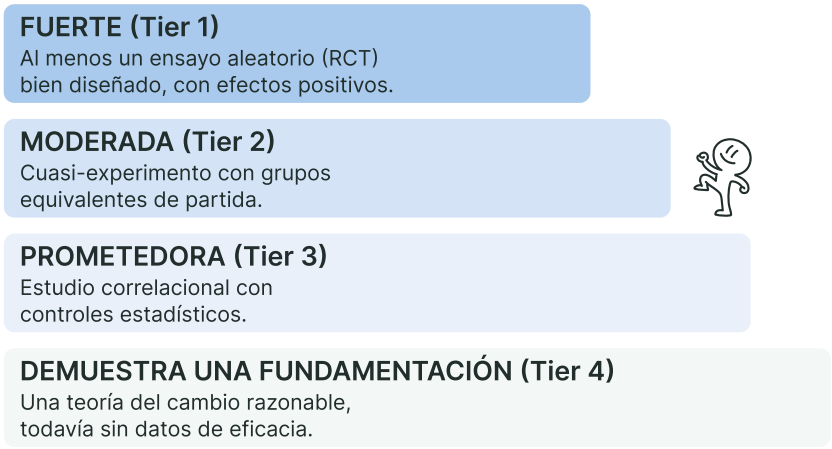

Figura 2. Los cuatro niveles de evidencia de la Every Student Succeeds Act (2015), ordenados de mayor a menor confianza causal.

Conviene tener cuidado con la interpretación. **Esta escala ordena la fuerza de la evidencia disponible sobre los programas, y no la calidad de los programas en sí.** Uno situado en un nivel bajo puede ser prometedor y merecer una evaluación más rigurosa. Uno con evidencia fuerte cuenta con estudios que permiten atribuirle sus resultados con mucha mayor confianza.

Y ni siquiera un buen experimento responde a todas las preguntas. Nos dice si encontramos una diferencia entre dos grupos comparables, y todavía tendremos que preguntarnos qué se ha medido, cuánto ha mejorado el alumnado y si podemos esperar algo parecido en nuestro propio contexto. Son las preguntas que abordaremos a continuación.

> La pregunta fundamental es por qué podemos creer que la diferencia encontrada se debe al programa. La respuesta depende, sobre todo, de cómo se haya construido la comparación.

3 | Antes de medir el impacto, hay que saber qué esperamos que cambie

Ya sabemos cómo construir una comparación que permita estimar el efecto de un programa. Nos falta todavía una pregunta anterior. ¿Qué esperamos que cambie si el programa funciona?

Imaginemos un nuevo programa de matemáticas que propone trabajar con más problemas abiertos. Si funciona, el alumnado tendrá más oportunidades de explicar cómo ha llegado a una solución, eso podría mejorar su capacidad de razonar y, con el tiempo, ese mejor razonamiento podría traducirse en un mejor desempeño en problemas que no ha practicado directamente.

Esta cadena de pasos es lo que se conoce como **teoría del cambio**, una explicación de cómo esperamos que una intervención produzca sus resultados. No hace falta que sea complicada. En esencia, nos obliga a responder a una pregunta muy sencilla. ¿Por qué pensamos que hacer *A* debería producir un cambio en *B*?

Del programa al aprendizaje

Sigamos con el mismo ejemplo. Un programa basado en que los docentes planteen más problemas abiertos recorrería cuatro etapas.

1. El programa se implementa. Los docentes reciben formación y empiezan a utilizar las nuevas actividades.

2. Cambia la práctica en el aula. Los alumnos trabajan efectivamente más problemas abiertos y tienen más oportunidades de explicar sus estrategias.

3. Cambia el aprendizaje. Los alumnos desarrollan mejores estrategias de razonamiento y resolución de problemas.

4. Cambian los resultados que nos interesan. Los alumnos obtienen mejores resultados en una prueba de resolución de problemas o en una medida más general de competencia matemática.

Cada flecha de esa cadena contiene una hipótesis, y cada hipótesis puede fallar. Quizá el programa aumente el número de problemas que se trabajan y deje intactas las estrategias que utiliza el alumnado. Quizá los alumnos desarrollen nuevas estrategias y no sepan utilizarlas cuando se enfrentan a problemas diferentes de los practicados. Quizá mejoren en resolución de problemas y ese cambio sea demasiado pequeño para observarse en una prueba general de matemáticas.

Por eso, cuando un estudio no encuentra efecto, tenemos a mano una pregunta más útil que «¿el programa funciona o no funciona?». Podemos preguntarnos en qué punto de la cadena dejó de producirse el cambio esperado. Puede que los docentes no hayan podido aplicar el programa como estaba previsto. Puede que las actividades se hayan utilizado sin cambiar realmente la experiencia de aprendizaje del alumnado. Puede que el aprendizaje haya cambiado y la prueba utilizada no lo detecte. O puede que la hipótesis inicial fuera sencillamente incorrecta, y hacer más problemas abiertos no produzca el tipo de aprendizaje que esperábamos.

Mirar la cadena completa nos ayuda a distinguir entre estas posibilidades. Y también cambia cómo interpretamos un resultado positivo. Si encontramos una mejora en la prueba final, podemos comprobar si observamos también los cambios intermedios que esperábamos. Cuando aparecen, la historia causal resulta más coherente. Cuando faltan, conviene ser más prudentes al interpretar el resultado o revisar nuestra explicación de cómo funciona el programa.

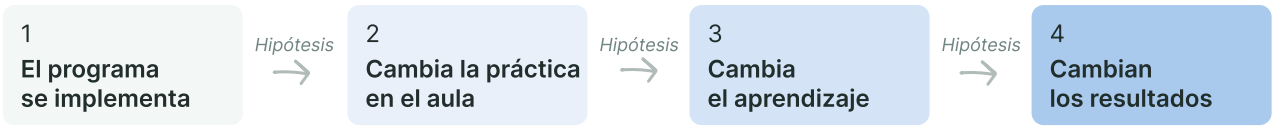

Figura 3. Una teoría del cambio conecta la intervención con los resultados que esperamos observar. Cada paso de la cadena plantea una hipótesis que puede comprobarse.

Una teoría del cambio explica por qué esperamos que un programa produzca determinados resultados. Nos permite pasar de «este programa funciona» a una pregunta mucho más útil: ¿qué debería cambiar, a través de qué mecanismo y bajo qué condiciones?

¿Qué significa «mejorar en matemáticas»?

La teoría del cambio nos obliga también a concretar qué entendemos por aprendizaje. Un programa puede estar diseñado para mejorar el cálculo, la comprensión conceptual, la resolución de problemas o la capacidad de razonar. También puede perseguir cambios en la confianza del alumnado, su actitud hacia las matemáticas o su disposición a enfrentarse a problemas difíciles. Todas estas cosas pueden ser relevantes, y **solo sabremos si un programa ha conseguido aquello que pretendía si medimos el resultado adecuado**.

Conviene, por tanto, distinguir entre distintos tipos de resultados.

- **Resultados próximos.** Los más directamente relacionados con lo que se ha enseñado, como dominar una estrategia concreta o resolver correctamente los tipos de ejercicios que se han practicado.

- **Transferencia.** Si el alumnado puede utilizar lo aprendido en situaciones nuevas, con problemas diferentes, formulaciones distintas o tareas que no ha practicado directamente.

- **Resultados amplios o duraderos.** Cambios más generales o que se mantienen en el tiempo, como una mejora global de la competencia matemática, una mayor confianza o efectos posteriores sobre la trayectoria académica.

Los tres pueden ser importantes y responden a preguntas diferentes. Un programa puede producir un efecto claro sobre un resultado próximo y mucho menor sobre una prueba más general. Eso puede significar que el aprendizaje todavía no se ha transferido, que la intervención actúa sobre una habilidad muy concreta o simplemente que estamos midiendo algo distinto de lo que el programa pretende cambiar.

La prueba forma parte de la pregunta

Elegir una prueba es, por tanto, mucho más que un detalle técnico posterior al diseño del estudio. **La prueba determina, en buena medida, qué afirmación podemos hacer sobre el programa.**

Imaginemos que un programa enseña una determinada estrategia para resolver problemas y que, al final del curso, el alumnado obtiene mejores resultados en ejercicios prácticamente idénticos a los que ha practicado. Podemos concluir que ha aprendido esa estrategia. Concluir que ha mejorado su capacidad general para resolver problemas sería mucho más arriesgado mientras no comprobemos qué ocurre cuando las tareas son nuevas.

Hay una razón adicional para prestar atención a esto. Las pruebas diseñadas por quienes desarrollan un programa suelen estar especialmente bien alineadas con aquello que se enseña, y eso tiende a hacer que el efecto parezca mayor. Cheung y Slavin (2016) revisaron 645 estudios educativos y encontraron que los efectos medidos con pruebas creadas por los propios investigadores eran aproximadamente el doble de grandes que los medidos con pruebas independientes. Los metaanálisis de intervenciones STEM encuentran el mismo patrón (Lynch et al., 2019).

Tiene sentido. Si durante meses practicamos un determinado tipo de ejercicio y después evaluamos precisamente ese tipo de ejercicio, estamos midiendo en parte lo cerca que está el aprendizaje de la práctica recibida. Una prueba externa y más amplia plantea una exigencia mayor. ¿Puede el alumnado utilizar lo aprendido cuando cambia la tarea?

Nada de esto convierte una prueba propia en mala herramienta. Si queremos saber si el alumnado ha aprendido exactamente un contenido o una estrategia que el programa pretende enseñar, una prueba propia puede ser la mejor opción disponible. Cada prueba permite hacer una afirmación diferente, y por eso conviene comprobar siempre quién diseñó la que se utilizó y hasta qué punto estaba alineada con el programa.

PREGUNTAS ABIERTAS

- ¿Qué pretendía cambiar el programa?
- ¿Qué mecanismo debería producir ese cambio?
- ¿Qué resultado se ha medido?
- ¿La prueba mide aquello que realmente queremos mejorar?
- ¿Es una prueba próxima a lo que se ha practicado o exige transferir lo aprendido?

Para interpretar un estudio de impacto conviene entender qué cambio pretendía producir el programa, qué mecanismo debía llevar hasta ese resultado y qué partes de esa cadena podemos observar. La teoría del cambio nos dice también dónde mirar cuando queramos saber si el resultado puede trasladarse a otra escuela, algo a lo que volveremos más adelante.

4 | ¿Cuánto ha cambiado y con qué precisión lo sabemos?

Hemos visto cómo construir una comparación que permita atribuir una diferencia al programa y cómo decidir qué resultado queremos medir. Supongamos ahora que hacemos el estudio y encontramos que los alumnos que utilizaron el programa obtuvieron mejores resultados. La siguiente pregunta es cuánto mejores.

El resultado de un estudio es siempre una estimación. Para interpretarla necesitamos responder a dos preguntas. Cuál es la magnitud de la diferencia y con qué precisión la hemos estimado. La primera nos la da el **tamaño del efecto**. La segunda, el **intervalo de confianza**. En los estudios suele aparecer también el **valor p**, que nos ayuda a valorar hasta qué punto los datos son compatibles con la ausencia de un efecto y guarda silencio sobre si la diferencia encontrada es grande o relevante desde el punto de vista educativo.

Tamaño del efecto. Cuánto de grande es la diferencia entre el grupo que usó el programa y el de comparación, medida en desviaciones estándar (SD, por sus siglas en inglés). Suele expresarse como la *d* de Cohen o la *g* de Hedges, dos maneras ligeramente distintas de calcular lo mismo, que llevan el nombre de los estadísticos que las propusieron.

Las pruebas no utilizan todas la misma escala. Una puede puntuar de 0 a 10, otra de 0 a 100 y otra utilizar una escala propia. Para poder comparar resultados entre estudios expresamos la diferencia en relación con la variabilidad de los resultados entre alumnos, que es justamente lo que mide la desviación estándar. Un efecto de 0,20 SD significa que la diferencia entre los dos grupos equivale a una quinta parte de esa variabilidad típica.

Dicho de otra manera, si ponemos a los veinticinco alumnos de una clase en fila, ordenados por resultados, un efecto de 0,20 SD desplazaría al alumno que estaba en medio unos dos puestos hacia arriba.

Cuánto mejora un alumno con un efecto de 0,20 SD

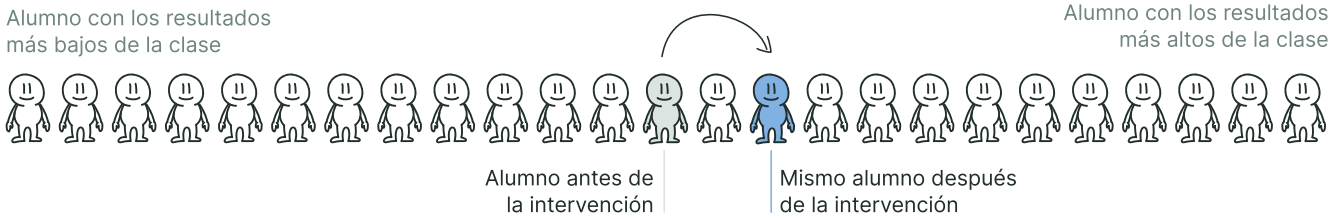

Figura 4. Los veinticinco alumnos de una clase ordenados por resultados. Un efecto de 0,20 SD desplaza al alumno que estaba en medio unos dos puestos.

Encontrar un efecto de 0,20 SD todavía no nos dice cuál es el efecto verdadero. Si repitiéramos el estudio con otras escuelas comparables, probablemente obtendríamos una cifra algo diferente. El intervalo de confianza expresa esa incertidumbre. Un efecto estimado de 0,20 SD con un intervalo de 0,15 a 0,25 es una estimación bastante precisa. El mismo efecto de 0,20 SD con un intervalo de −0,05 a 0,45 es mucho más incierto, porque los datos resultan compatibles tanto con un efecto prácticamente nulo como con uno muy grande. El problema es más frecuente de lo que parece: una revisión de los grandes ensayos educativos financiados en Estados Unidos y el Reino Unido encontró que muchos producen estimaciones demasiado imprecisas para resultar informativas (Lortie-Forgues e Inglis, 2019). Por eso conviene leer siempre el tamaño del efecto junto a su intervalo de confianza.

Un tamaño del efecto se interpreta junto a su intervalo de confianza. Necesitamos saber tanto la magnitud estimada como la incertidumbre que la rodea.

¿0,20 SD es mucho o poco? Depende de con qué lo comparemos

Un tamaño del efecto necesita una referencia para poder interpretarse. Necesitamos saber qué magnitudes suelen encontrarse en estudios parecidos.

Kraft (2020) propuso una referencia ajustada a la realidad escolar, construida a partir de la distribución de efectos obtenidos en ensayos aleatorizados en educación, y la actualizó tres años más tarde con una muestra mucho mayor. Reunió 3.426 tamaños de efecto procedentes de 973 ensayos aleatorizados y comprobó que los efectos suelen ser modestos. Más de un tercio estaba por debajo de 0,05 SD, la mediana era de 0,10 SD y solo el 10 % superaba 0,57 SD (Kraft, 2023). A partir de esta distribución propuso una escala específica para intervenciones educativas: menos de 0,05 SD es un efecto pequeño, entre 0,05 y 0,20 SD un efecto medio, y a partir de 0,20 SD un efecto grande. Una distribución muy similar apareció en 138 ensayos realizados en países de ingresos bajos y medios, también con una mediana de 0,10 SD (Evans y Yuan, 2022).

La referencia que utilicemos importa mucho. En educación se han utilizado también otros marcos, como el de Cohen (1988), desarrollado para las ciencias del comportamiento en general, o el conocido umbral de 0,40 de Hattie (2009). Ambos han tenido mucha influencia y se construyeron a partir de materiales muy diversos, sin limitarse a la distribución de efectos obtenidos en experimentos rigurosos realizados en contextos educativos reales. Aplicarlos directamente nos llevaría a descartar como pequeños unos efectos que, comparados con los que suelen conseguir las intervenciones educativas, son claramente relevantes.

Por eso un efecto de 0,20 SD puede parecer modesto en términos absolutos y ser, al mismo tiempo, un resultado importante en educación. Junto a la pregunta de cuánto mide el efecto, conviene hacerse otra. ¿Cómo se compara con lo que consiguen intervenciones similares?

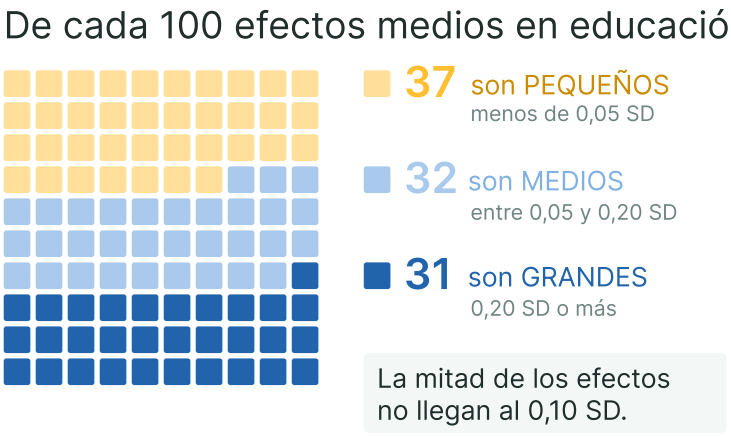

Figura 5. Distribución de 3.426 tamaños de efecto procedentes de 973 ensayos aleatorizados en educación, con la mediana y los tres tramos propuestos por Kraft. Elaboración propia a partir de Kraft (2020, 2023).

Hay una segunda cautela importante. Un mismo tamaño del efecto puede representar una mejora más o menos relevante dependiendo del curso en el que se mida.

> «Un efecto de 0,20 SD puede parecer modesto en términos absolutos y ser, al mismo tiempo, un resultado importante en educación.»

Hill, Bloom, Black y Lipsey (2008) estimaron cuánto suelen progresar los alumnos en pruebas estandarizadas a lo largo de un curso escolar y mostraron que ese progreso varía mucho según la edad. En los primeros cursos de primaria, el crecimiento observado puede acercarse a una desviación estándar completa por año. Hacia quinto o sexto se sitúa alrededor de 0,40 SD, y en los últimos cursos de secundaria queda por debajo de 0,25 SD. En otras palabras, en los primeros cursos hay mucho margen de mejora, y a medida que los estudiantes son más competentes cada vez cuesta más mostrar progreso.

Tenemos así una segunda referencia, junto a los efectos de otras intervenciones, para interpretar un tamaño del efecto.

Capacidad de mejora de un alumno a medida que avanza

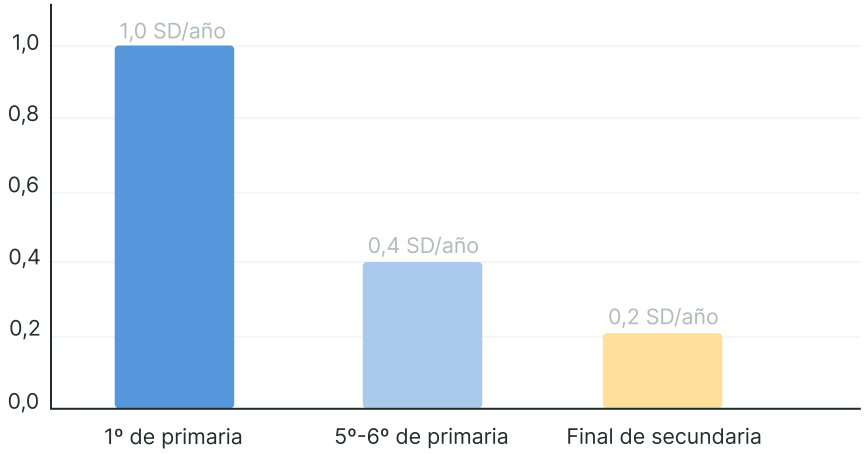

Figura 6. Crecimiento anual típico según el curso, a partir de Hill et al. (2008).

Un efecto de 0,10 SD, que es el orden de magnitud habitual de los materiales curriculares de matemáticas (Slavin y Lake, 2008; Pellegrini et al., 2021), representa aproximadamente un mes del progreso que suele observarse durante un curso en los primeros años de primaria. En los últimos cursos de secundaria, donde el crecimiento anual observado es mucho menor, esos mismos 0,10 SD representan una fracción considerablemente mayor del progreso de un curso.

A veces estas comparaciones se expresan como «meses de aprendizaje». Pueden ser útiles para hacer más comprensible un tamaño del efecto y conviene interpretarlas con mucha cautela. Un estudio observa una diferencia en resultados y la compara con el crecimiento que suele producirse durante un curso; hablar de meses es una traducción posterior, que depende del curso, de la prueba y de la referencia utilizada.

Meses que tarda un alumno en mejorar 0,10 SD

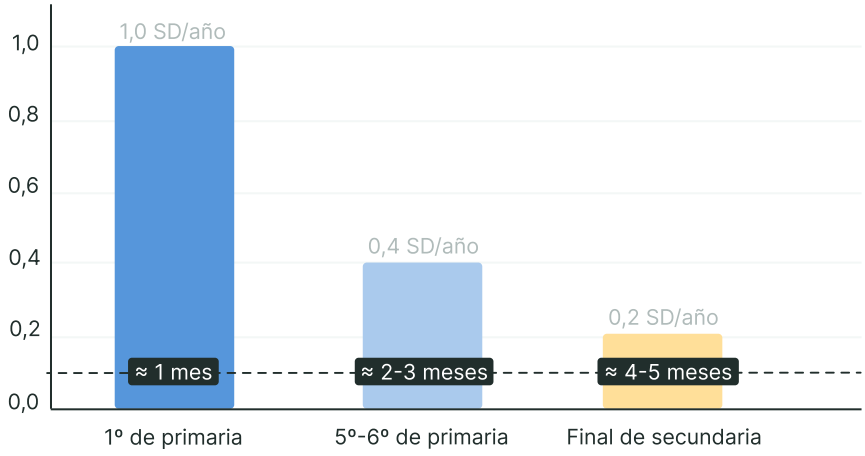

Figura 7. Aproximación de cuántos meses del progreso anual típico representa un efecto de 0,10 SD, según el curso, a partir de Hill et al. (2008). Estas equivalencias son orientativas, porque los estudios de impacto no miden directamente meses de aprendizaje.

El resultado de un estudio es una estimación. Para interpretarlo necesitamos saber cuánto cambia el resultado —el tamaño del efecto— y con qué precisión lo hemos estimado —el intervalo de confianza —. Después necesitamos poner esa magnitud en contexto, comparándola con los efectos que suelen conseguir intervenciones similares y con el progreso habitual del alumnado a esa edad. **La misma cifra puede parecernos pequeña o grande dependiendo de la referencia que utilicemos.**

5 | ¿Qué nos dice un estudio sobre un programa?

Hasta ahora hemos hablado de efectos de 0,10 o 0,20 SD como si fueran una característica del programa. Conviene evitar esa interpretación. Un estudio estima lo que ocurrió cuando un programa concreto se aplicó de una determinada manera, a unos alumnos determinados, frente a una alternativa concreta y utilizando una medida concreta. Si cambian estas condiciones, el efecto que observamos también puede cambiar. De hecho, cuando un metaanálisis reúne 191 experimentos sobre intervenciones matemáticas, buena parte de la variación entre resultados se explica por estas diferencias de contexto y de diseño (Williams et al., 2022).

Para saber qué nos dice realmente un estudio tenemos que leer conjuntamente las piezas que hemos ido viendo, porque la calidad de la evidencia depende de todas ellas a la vez. Un efecto grande resulta poco convincente si los grupos no eran comparables. Un estudio muy bien diseñado puede responder a una pregunta poco relevante para nosotros si mide únicamente tareas muy próximas al programa. Y un efecto positivo puede seguir siendo incierto si el intervalo de confianza es muy amplio.

Incluso cuando todo encaja, dos estudios rigurosos sobre el mismo programa pueden encontrar efectos distintos sin que ninguno de los dos esté equivocado. Cuatro preguntas ayudan a entender por qué.

¿Con qué se midió?

Como vimos en la sección 3, la prueba utilizada determina en parte qué entendemos por «funcionar». Los estudios sobre intervenciones educativas suelen encontrar efectos mayores en pruebas diseñadas o estrechamente alineadas con el programa que en medidas más independientes (Cheung y Slavin, 2016; Lynch et al., 2019). Una prueba próxima puede mostrarnos si se ha aprendido aquello que el programa enseñaba directamente, y una medida más alejada, si ese aprendizaje se transfiere a tareas nuevas. La pregunta importante es qué aprendizaje representa el efecto que estamos observando.

¿Con qué se comparó?

El tamaño del efecto depende también de la alternativa con la que se compara el programa. Comparar una intervención con una práctica poco estructurada es muy distinto de compararla con otro programa de matemáticas bien implementado. En ambos casos el estudio puede estar perfectamente diseñado y responder a preguntas diferentes. Conviene, por tanto, saber qué habría recibido el grupo de comparación si no hubiera participado en el programa. **Cuanto mejor sea la alternativa disponible, más exigente será demostrar una mejora adicional.**

¿Cuándo se midió?

El momento de la medición también importa. Un programa puede producir una mejora al terminar la intervención que se reduzca, desaparezca o, en algunos casos, aumente con el tiempo. La revisión más completa sobre esta cuestión concluye que una pérdida parcial del efecto es lo habitual, y que en muchas intervenciones los estudios de seguimiento recuperan solo una fracción del efecto inicial (Bailey et al., 2020). Si lo que nos interesa es saber si el aprendizaje perdura, necesitamos medidas de seguimiento. Un efecto observado justo al finalizar el programa nos dice que hubo una diferencia en ese momento, y cuánto durará es una pregunta aparte.

¿En qué condiciones se aplicó?

Finalmente, necesitamos entender qué hizo falta para obtener el resultado. Un programa evaluado en unas pocas escuelas, con docentes especialmente motivados, formación intensiva y apoyo continuo del equipo que lo desarrolló puede obtener resultados distintos cuando se extiende a cientos de centros en

condiciones ordinarias. Es un patrón bien documentado. En el acompañamiento docente, por ejemplo, los efectos de los programas grandes son solo una fracción de los que obtienen los pilotos pequeños (Kraft, Blazar y Hogan, 2018).

Esto no convierte un estudio pequeño en un mal estudio. Nos dice qué condiciones acompañaron al efecto y nos obliga a preguntarnos cuáles de ellas eran necesarias para producirlo. Como vimos al hablar de la teoría del cambio, la **implementación** es uno de los eslabones que conectan el programa con el aprendizaje, y un buen programa puede quedarse sin efecto si no llega a transformar de forma consistente lo que ocurre en el aula (Durlak y DuPre, 2008). Combinar materiales de calidad, formación docente y acompañamiento centrado en la práctica suele funcionar mejor que aplicar estos elementos por separado (Lynch et al., 2019; Piper et al., 2018). Por eso resulta especialmente informativa la evidencia que muestra que los efectos se mantienen cuando un programa pasa de condiciones muy controladas a una implementación amplia y parecida a la práctica habitual.

Es también la razón por la que la comisión del National Research Council que revisó la evaluación de programas curriculares de matemáticas concluyó que ningún estudio aislado basta, y propuso cruzar tres miradas complementarias: el impacto sobre los resultados, la implementación en las aulas y la calidad matemática del contenido (Confrey y Stohl, 2004).

¿Quién ha realizado el estudio?

Conviene saber también quién ha realizado y financiado la evaluación. Las revisiones de la literatura educativa encuentran que los estudios realizados o encargados por quienes desarrollan una intervención tienden a reportar efectos mayores que las evaluaciones independientes (Wolf et al., 2020). Los estudios pequeños y los que llegan a publicarse también tienden a mostrar efectos más grandes que los estudios amplios y que los que se quedan en un cajón (Cheung y Slavin, 2016; Slavin y Smith, 2009; Franco, Malhotra y Simonovits, 2014).

Nada de esto invalida un estudio financiado por el desarrollador, cuyo diseño debe evaluarse con los mismos criterios que cualquier otro. Es una razón más para valorar especialmente que los resultados hayan sido replicados por equipos independientes.

¿Qué significa este resultado para nuestra escuela?

Un estudio riguroso nos dice qué ocurrió en unas condiciones determinadas. Trasladar ese resultado a otra escuela exige un paso más, y ese paso consiste en algo distinto de buscar estudios realizados en centros idénticos al nuestro. La pregunta importante es si se dan en nuestro contexto las condiciones que eran necesarias para que el programa produjera el cambio.

Aquí volvemos a la teoría del cambio. Si el programa necesitó determinada formación docente, una cantidad concreta de tiempo en el aula o ciertas prácticas para producir el aprendizaje, debemos preguntarnos si esas condiciones pueden reproducirse en nuestro centro. Cuanto más dependa el resultado de condiciones que nosotros no podemos reproducir, menos informativo será ese efecto para anticipar lo que ocurriría aquí.

Por la misma razón, más que buscar «el estudio» que demuestra que un programa funciona, conviene mirar el conjunto de la evidencia disponible. Un único estudio puede proporcionar evidencia causal muy convincente sobre lo que ocurrió en un contexto concreto. Nuestra confianza aumenta cuando encontramos resultados consistentes en varios estudios rigurosos y cuando esos resultados se mantienen al cambiar las poblaciones, las escuelas, las medidas o las condiciones de implementación. La replicación es escasa en investigación educativa, y precisamente por eso resulta tan valiosa cuando aparece (Makel y Plucker, 2014).

Si un efecto aparece únicamente con una prueba muy próxima, en un estudio pequeño realizado con mucho apoyo y justo al terminar la intervención, sabemos algo importante y limitado. Si también aparece con medidas más alejadas, frente a alternativas exigentes, en implementaciones a mayor escala y en contextos distintos, tenemos más razones para pensar que el resultado no depende de una circunstancia muy particular.

Un estudio nos proporciona una estimación de lo que ocurrió en unas condiciones determinadas. Para interpretarla necesitamos leer conjuntamente la calidad de la comparación, qué se midió, la magnitud y precisión del efecto, con qué se comparó el programa, cuándo se midieron los resultados y en qué condiciones se aplicó. La pregunta final es qué nos permite concluir este estudio y qué razones tenemos para esperar un resultado parecido en nuestro contexto. Cuanto más consistente sea la evidencia entre estudios rigurosos, medidas, poblaciones y condiciones diferentes, mayor será nuestra confianza en que el resultado no depende de una circunstancia particular.

6 | Una guía para leer la evidencia

A lo largo de este capítulo hemos visto que decidir si un programa «funciona» requiere bastante más que encontrar un estudio con resultados positivos. **Estas ocho preguntas resumen una forma práctica de leer cualquier estudio sobre un programa educativo.**

1. ¿Qué se está comparando? Qué reciben los alumnos que utilizan el programa y qué recibe el grupo de comparación. El efecto siempre es una diferencia respecto a una alternativa concreta.

2. ¿Podemos atribuir esa diferencia al programa? Si hubo asignación aleatoria y, cuando no la hubo, qué estrategia permite pensar que los grupos eran comparables antes de empezar.

3. ¿Qué esperaba cambiar el programa y cómo? Cuál es su teoría del cambio y qué pasos intermedios podemos observar.

4. ¿Qué se midió realmente? Si la medida captura lo que el programa enseñaba directamente o exige transferir lo aprendido a tareas nuevas, y si es el resultado que nos interesa.

5. ¿Cuánto cambió y con qué precisión lo sabemos? El tamaño del efecto, su intervalo de confianza y cómo se comparan con los de intervenciones similares y con el progreso habitual del alumnado a esa edad.

6. ¿En qué condiciones se obtuvo el resultado? Frente a qué alternativa, en qué momento se midió, con qué formación, tiempo y apoyo contaron las escuelas.

7. ¿Qué sabemos más allá de este estudio? Si existen otros estudios rigurosos e independientes sobre el programa y si encuentran resultados consistentes al cambiar medidas, poblaciones y escuelas.

8. ¿Qué significa esta evidencia para nuestro contexto? Si podemos reproducir las condiciones que parecen necesarias para que el programa produzca el cambio.

La pregunta útil es qué demuestra un estudio, con qué grado de confianza y hasta qué punto esa evidencia resulta relevante para la decisión que queremos tomar.

7 | Preguntas frecuentes

«Si todos los estudios tienen limitaciones, ¿de qué me sirve la evidencia?»

La evidencia no sustituye el criterio profesional; lo hace más informado. En la decisión también importan las necesidades del alumnado, los recursos disponibles, el coste y la capacidad del centro para implementar el programa. Un efecto moderado puede ser muy valioso si el programa es asequible y fácil de aplicar, y mucho menos interesante si exige una inversión desproporcionada. Lo que sí puede hacer la evidencia es ayudarnos a anticipar qué resultados son razonables, comparar alternativas, detectar promesas poco creíbles y mostrarnos qué condiciones parecen necesarias para que un programa produzca determinados resultados.

«Llevamos tres años con el mismo programa y los resultados han subido claramente. ¿Eso no cuenta?»

Cuenta, y es información valiosa sobre lo que está ocurriendo en vuestra escuela. Lo difícil es saber cuánto de esa mejora ha causado el programa. Durante esos tres años pueden haber cambiado el equipo docente, la formación recibida, las horas dedicadas a matemáticas o las características del alumnado. Conviene mirar también si la mejora es específica de vuestro centro o si acompaña a una tendencia general de la zona. **«Nuestros resultados han mejorado desde que utilizamos el programa» y «el programa ha causado esa mejora» son afirmaciones distintas**, y la primera puede ser perfectamente cierta mucho antes de que tengamos evidencia suficiente para sostener la segunda.

«Mi proveedor me enseña un estudio con un efecto de 0,45 SD. ¿Debería alegrarme o desconfiar?»

Las dos cosas, y en ese orden. Un efecto de 0,45 SD está más de cuatro veces por encima de la mediana de los ensayos educativos y más del doble del umbral a partir del cual Kraft ya considera un efecto grande. Si la estimación es creíble, sería un resultado excepcional. Precisamente por eso merece la pena entender de dónde sale la cifra. ¿Con qué prueba se midió? ¿Frente a qué alternativa? ¿Cuál es el intervalo de confianza? ¿Se midió también algún tiempo después? ¿Se evaluó en unas pocas escuelas con mucho apoyo o en condiciones parecidas a las de una implementación habitual? ¿Ha encontrado alguien más algo parecido? **Un efecto excepcional es una buena razón para leer el estudio con especial atención.**

«¿Cuánto debería fiarme de un sello de evidencia?»

Un sello indica que alguien ha revisado el estudio con criterios previamente establecidos, y eso es muy útil. Se parece al certificado de la ITV de un coche, que acredita haber superado una revisión exigente sin garantizar que el coche funcione igual en cualquier carretera. Nos informa sobre la evidencia disponible y su calidad, y seguimos teniendo que preguntarnos qué se midió, frente a qué se comparó el programa y en qué condiciones se obtuvo el resultado. Funciona también al revés. **Que un programa no tenga sello puede significar simplemente que opera donde no existe ningún organismo que los emita**, o que nadie ha presentado sus estudios a revisión. En ese caso tendremos que acudir directamente a los estudios y aplicarles las preguntas de este capítulo.

«¿Quién revisa los estudios y me puede ayudar a saber si la evidencia es buena?»

Existen organismos especializados que revisan estudios educativos con criterios comunes y sintetizan sus resultados. En Estados Unidos, el What Works Clearinghouse revisa la calidad metodológica de los estudios (What Works Clearinghouse, 2022) y Evidence for ESSA organiza la evidencia sobre programas según los niveles establecidos por la legislación. EdReports valora la calidad del currículo y su alineación con los estándares, sin entrar a evaluar el impacto. En Europa, el referente más cercano es la Education Endowment Foundation británica, que financia evaluaciones rigurosas y sintetiza evidencia educativa. También podemos acudir a estudios publicados en revistas con revisión por pares, donde otros especialistas examinan los métodos, los análisis y las conclusiones antes de la publicación. Superar ese proceso ofrece cierta garantía de que los resultados se interpretan correctamente y de que las limitaciones se explican con transparencia. La fuerza causal de un estudio sigue dependiendo, en cambio, de cómo se haya construido la comparación y de qué pregunta permita responder el diseño. Los organismos de síntesis, los sellos y la revisión por pares nos ayudan a orientarnos, y ninguno de ellos sustituye las preguntas fundamentales de este capítulo.

8 | Para terminar

La pregunta que abría el capítulo, «¿funciona?», tiene una versión mucho más útil, y es la misma que planteaba hace veinte años la comisión del National Research Council (Confrey y Stohl, 2004). Para quién funciona, en qué condiciones y por qué. Un buen estudio de impacto responde a la primera parte con una comparación creíble y una estimación honesta de su magnitud. Las otras dos las responde la teoría del cambio, y son las que nos permiten anticipar qué ocurrirá cuando el programa entre en nuestras aulas.

La pregunta útil es qué tiene que pasar para que este programa funcione aquí, con estos docentes y este alumnado.

En el capítulo 2 aplicaremos estas preguntas a la evidencia disponible sobre programas curriculares de matemáticas.

Para profundizar

Marcadas con ▸ las dos puertas de entrada para quien quiera seguir por su cuenta sin ser investigador.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., y Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest, 21*(2), 55–97.

Cheung, A. C. K., y Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.ª ed.). Lawrence Erlbaum Associates.

Confrey, J., y Stohl, V. (eds.) (2004). *On evaluating curricular effectiveness: Judging the quality of K–12 mathematics evaluations*. National Research Council, National Academies Press.

Durlak, J. A., y DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350.

Evans, D. K., y Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis, 44*(3), 532–540.

Every Student Succeeds Act (2015). Pub. L. No. 114-95, 129 Stat. 1802. Estados Unidos.

Franco, A., Malhotra, N., y Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505.

▸ Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., y Vermeersch, C. M. J. (2016). *Impact evaluation in practice* (2.ª ed.). Banco Interamericano de Desarrollo y Banco Mundial.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hill, C. J., Bloom, H. S., Black, A. R., y Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

▸ Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253.

Kraft, M. A. (2023). The effect size benchmark that matters most: Education interventions often fail. *Educational Researcher, 52*(3), 183–187.

Kraft, M. A., Blazar, D., y Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.

Lortie-Forgues, H., y Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166.

Lynch, K., Hill, H. C., Gonzalez, K. E., y Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–303.

Makel, M. C., y Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304–316.

Pellegrini, M., Lake, C., Neitzel, A., y Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open, 7*(1), 1–29.

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., y King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development, 106*, 324–336.
Slavin, R. E., y Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427–515.

Slavin, R. E., y Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500–506.

What Works Clearinghouse (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. Institute of Education Sciences, U.S. Department of Education.

Williams, R., Citkowicz, M., Miller, D. I., Lindsay, J., y Walters, K. (2022). Heterogeneity in mathematics intervention effects: Evidence from a meta-analysis of 191 randomized experiments. *Journal of Research on Educational Effectiveness, 15*(3), 584–634.

Wolf, R., Morrison, J., Inns, A., Slavin, R., y Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness, 13*(2), 428–447.

Financiación y declaración de intereses. *Esta publicación ha sido financiada por Innovamat Education y forma parte de una serie de divulgación científica sobre educación promovida por la entidad. Marc Colomer y Carles Aulés forman parte del equipo de Innovamat. Caterina Calsamiglia participa como autora externa, aportando su experiencia en evaluación de impacto educativo. Las afiliaciones y la financiación se hacen explícitas como parte del compromiso de transparencia de la publicación. El contenido y las conclusiones son responsabilidad de los autores.*

Agradecimientos. *Queremos agradecer a las personas que han revisado este capítulo y han contribuido a mejorar su calidad: Ángela Perez, Albert Vilalta, Anna Llobet, Claudia Portas y Aida Sanuy. También queremos agradecer el trabajo de maquetación de Marta Soler y su contribución en los gráficos.*