Innovamat

¿Qué impacto tienen los programas de matemáticas en el aprendizaje?

Cuánta evidencia hay, qué mejora es realista esperar y qué condiciones la multiplican.

Carles Aulés
Especialista de Investigación del Innovamat Research Lab

Marc Colomer
Director de Investigación del Innovamat Research Lab

Innovamat Research Series

Resumen

¿Qué sabemos sobre el impacto de los materiales y programas de matemáticas en el aprendizaje? La investigación acumulada desde comienzos de este siglo ofrece una conclusión general: **los materiales curriculares pueden mejorar los resultados, pero sus efectos suelen ser moderados y dependen en gran medida de cómo se implementan**. Además, todavía sabemos mucho menos de lo deseable sobre qué programas funcionan, para quién y en qué condiciones. La revisión de la evidencia disponible permite extraer las siguientes conclusiones:

- **La evidencia rigurosa sigue siendo escasa.** La mayor parte procede de Estados Unidos y, aun allí, muchos programas ampliamente utilizados nunca han sido evaluados con diseños capaces de estimar de forma convincente su impacto.

- **Obtener evidencia causal es costoso y difícil.** Una evaluación rigurosa puede requerir varios años, miles de estudiantes, numerosas escuelas, financiación considerable y centros dispuestos a aceptar que la asignación de los programas se determine de forma aleatoria.

- **Cuando se detectan mejoras, suelen ser moderadas, pero pueden ser relevantes a gran escala.** Los meta-análisis sitúan la ventaja media de sustituir unos materiales curriculares por otros alrededor de +0,05 a +0,10 desviaciones estándar (SD). Entre los estudios revisados por *What Works Clearinghouse* y *Evidence for ESSA*, la mediana es de +0,13 SD. Son efectos modestos, pero la comparación no es entre recibir o no recibir enseñanza: el grupo de control continúa utilizando sus materiales habituales. Por tanto, estas cifras representan la mejora adicional atribuible a sustituir una alternativa por otra. Si ambas tienen costes similares, incluso una mejora moderada puede resultar importante cuando alcanza a todos los alumnos.

- **Un mismo programa puede producir resultados distintos según dónde, cómo y contra qué se evalúe.** El impacto observado depende de la calidad de la alternativa utilizada por el grupo de comparación, de la prueba empleada para medir el aprendizaje, del tiempo transcurrido desde la intervención y de la calidad de la implementación. Las evaluaciones muy alineadas con el contenido del programa suelen encontrar efectos mayores que las pruebas externas, y parte de las mejoras puede disminuir con el tiempo.

- **Los materiales funcionan mejor cuando van acompañados de formación docente vinculada a la práctica.** Los estudios sintetizados por Pellegrini et al. (2021) y Lynch et al. (2019) encuentran medidas de efecto superiores cuando la intervención combina materiales curriculares con acompañamiento y desarrollo profesional (la medida del efecto crece entre +0.08 y 0.10 SD). La evidencia sugiere que entregar buenos materiales no basta: ayudar al profesorado a utilizarlos bien aumenta su impacto.

- **El equilibrio entre el coste y el beneficio.** Intervenciones intensivas como la tutoría o el trabajo en pequeños grupos obtienen, en promedio, efectos altos. Algo parecido pasa con el *coaching*. Estas intervenciones requieren, sin embargo, muchos más recursos humanos y son difíciles de extender. Los materiales curriculares acompañados de formación producen efectos menores por alumno, pero pueden llegar a prácticamente todas las aulas y hacerlo con un coste adicional relativamente bajo.

Cuando buscamos recursos, materiales y programas para apoyar la enseñanza y el aprendizaje de las matemáticas, la oferta resulta abrumadora. Hay libros de texto y otros materiales curriculares, aplicaciones digitales complementarias, herramientas de evaluación, servicios de formación docente y programas integrales que combinan materiales, formación y acompañamiento. Son propuestas muy distintas en su enfoque y alcance. Ante esta variedad, la pregunta central es qué diferencia puede producir cada opción en el aprendizaje del alumnado. ¿Importa realmente escoger un recurso en lugar de otro? ¿Qué mejora cabe esperar de una propuesta de mayor calidad?

Antes de nada, vale la pena precisar el vocabulario que utilizamos a lo largo del capítulo.

- Con **materiales curriculares o programa de matemáticas**, nos referimos a un conjunto estructurado de recursos educativos diseñados para apoyar la enseñanza de los contenidos establecidos en el currículo oficial. Pueden incluir libros de texto, guías para docentes, materiales manipulativos, herramientas digitales o de evaluación, entre otros componentes. Para distinguir mejor entre propuestas de distinto alcance, a lo largo del capítulo hablaremos generalmente de **materiales curriculares** cuando estos se ofrecen solos o con un apoyo a la implementación mínimo, y de **programa** cuando los materiales se acompañan de un apoyo más sustantivo a la implementación, como formación específica, asesoramiento pedagógico y/o *coaching*.

- Con **recursos suplementarios**, nos referimos a materiales estructurados que apoyan la enseñanza de determinados contenidos o habilidades, pero que no pretenden cubrir el currículo de manera integral: aplicaciones digitales, cuadernos de práctica, herramientas de evaluación, u otros materiales específicos.

- Con **formación docente**, nos referimos a las actividades orientadas al desarrollo de los conocimientos, las habilidades y, en general, la capacidad profesional de los docentes.

- **Intervención**: en su sentido más amplio, nos referimos a cualquier propuesta deliberada para mejorar la enseñanza o el aprendizaje de las matemáticas. Pero cuando hablamos de **intervención focalizada**, nos referimos a algo más concreto: la instrucción en grupos pequeños, complementaria a la clase y normalmente dirigida a alumnado con dificultades.

En cualquiera de estos casos, interpretar un estudio de impacto sobre una intervención educativa exige una mirada crítica, como explicamos en el capítulo 1 «¿Cómo sabemos si un programa de matemáticas funciona?». Debemos preguntarnos qué tipo de estudio se realizó, con qué alternativa se comparó, qué resultados se midieron, en qué momento y en qué contexto. Sin embargo, valorar el rigor de un estudio no basta. También necesitamos un punto de referencia que nos permita interpretar la magnitud del resultado: incluso cuando un efecto está bien estimado, ¿es grande o pequeño? En términos prácticos, **¿qué mejora es razonable esperar al adoptar un determinado programa de matemáticas?**

Para responder a estas preguntas, revisamos en profundidad la evidencia disponible. Antes de presentar los resultados, es útil delimitar su alcance. En nuestra revisión, Estados Unidos es el contexto en el que encontramos más evaluaciones del impacto de programas de matemáticas. En otros países y regiones existen numerosos estudios sobre prácticas de aula, trayectorias de aprendizaje o materiales concretos, pero son menos frecuentes las evaluaciones de un programa de matemáticas integral. Por ello, este capítulo se apoya principalmente en fuentes estadounidenses. Sus resultados pueden ofrecer un punto de partida útil para otros sistemas educativos, pero no deben trasladarse de manera automática: su relevancia dependerá, entre otros factores, del currículo oficial, la organización escolar, la formación docente y las características del alumnado de cada contexto.

Nuestro análisis combina tres fuentes de información. En primer lugar, examinamos los principales meta-análisis de la literatura científica. Un **meta-análisis** sintetiza de manera sistemática los resultados de varios estudios sobre una misma pregunta y permite estimar tanto el efecto medio como su variación. No obstante, sus conclusiones son tan sólidas como los estudios que incluye. En segundo lugar, codificamos 85 estudios recogidos, con fecha de corte en inicios de agosto de 2026, en *What Works Clearinghouse* y Evidence for ESSA, dos repositorios estadounidenses que revisan la evidencia sobre intervenciones educativas mediante criterios explícitos. En tercer lugar, recopilamos 229 documentos que seis grandes editoriales estadounidenses presentaban en sus sitios web como evidencia de sus productos. Estos documentos no son necesariamente equivalentes a estudios de impacto independientes; precisamente por eso, analizamos qué tipo de evidencia aportaba cada uno y cuál era su solidez.

De esta revisión se desprenden **seis conclusiones**, que desarrollamos a lo largo del capítulo. La evidencia de impacto rigurosa es la excepción, incluso en Estados Unidos. Generarla es lento, caro y difícil. Cuando existe, el efecto suele ser modesto, aunque puede ser una inversión eficiente. Los mismos materiales curriculares no siempre tienen el mismo impacto, porque dependen de con qué prueba, cuándo y en qué condiciones se evalúan. El impacto no depende solo de la calidad de los materiales, sino también de lo que viaja con ellos: la formación, el acompañamiento docente y las condiciones que permiten llevarlos adecuadamente al aula. Y, comparado con otras palancas como la tutoría o el *coaching*, los materiales curriculares con formación tienen un efecto menor por alumno pero mucho mayor alcance: hay una tensión entre impacto y escala.

Al final de este capítulo

sabrás qué proporción de los programas de matemáticas cuenta con evidencia de impacto sólida y por qué es tan difícil generarla; qué mejora es razonable esperar de un cambio de programa de matemáticas, por lo general modesta, aunque potencialmente relevante; por qué unos mismos materiales curriculares no siempre obtienen el mismo resultado; qué condiciones pueden amplificar su efecto, especialmente la formación y el acompañamiento docente; y qué tensión existe entre el impacto y la escala de las distintas palancas que puede activar una escuela.

Tamaño del efecto. La diferencia entre el resultado medio del grupo que recibió un programa y el del grupo de comparación, expresada en desviaciones estándar —SD, por sus siglas en inglés—. En una evaluación causal bien diseñada, esa diferencia permite estimar cuánto cambió el resultado debido al programa.

En el artículo «¿Cómo sabemos si un programa de matemáticas funciona?» explicamos con más detalle cómo interpretar un tamaño del efecto a nivel cualitativo, pero aquí basta con recordar dos ideas.

La primera es que necesitamos una referencia realista, pero no una regla universal. En la revisión ampliada de Kraft, la mediana de 3.426 estimaciones procedentes de 973 ensayos aleatorizados fue de 0,10 SD. Como punto de partida para estudios causales de intervenciones educativas evaluadas mediante pruebas estandarizadas de rendimiento, Kraft propone considerar pequeños los efectos inferiores a 0,05 SD; medios, los situados entre 0,05 y menos de 0,20 SD; y grandes, los de 0,20 SD o más. Estas categorías ayudan a orientarnos, pero no sustituyen el análisis del diseño, la medida utilizada, la población, el coste o la posibilidad de aplicar el programa a escala (Kraft, 2020, 2023).

La segunda es que una misma cifra en SD no representa la misma proporción del progreso anual típico en todos los cursos. En los datos históricos de pruebas estandarizadas estadounidenses analizados por Hill et al., las ganancias anuales medidas en SD son mayores en los primeros cursos y menores en secundaria. Si, como aproximación, dividimos ese progreso anual entre nueve meses, 0,10 SD equivale a cerca de un mes del progreso observado en los primeros cursos de primaria y a alrededor de cuatro meses entre 8º y 10º curso. Esta equivalencia sirve para construir intuición, pero no debe interpretarse como una medida literal de meses de aprendizaje: depende del curso, de la prueba y de varios supuestos sobre cómo se distribuye el progreso durante el año (Hill et al., 2008; Kraft, 2020).

1 | La evidencia de impacto rigurosa es la excepción, no la norma

¿Cómo de común es la evidencia de impacto rigurosa?

En 2023, Bellwether examinó la evidencia disponible sobre 59 programas de matemáticas de editoriales estadounidenses. Antes de seguir, intenta responder a esta pregunta: ¿para cuántos de esos programas crees que encontraron al menos un estudio de eficacia que cumpliera unos estándares mínimos de calidad científica?

La respuesta es 17. Para los otros 42 —aproximadamente siete de cada diez—, Bellwether no localizó ningún estudio que satisficiera todos sus requisitos. Entre otros criterios, el estudio debía haberse realizado en Estados Unidos, centrarse en alumnado de K–9, incluir al menos 250 participantes y haberse publicado en alguno de los formatos admitidos por la revisión. Cabe resaltar que estos 59 programas no se escogieron de manera aleatoria. Se trata de 59 programas que EdReports —una institución independiente y sin ánimo de lucro que evalúa y publica revisiones por expertos sobre materiales instructivos— identificó como de calidad media o alta. Es decir, se trata de programas que a nivel de contenido curricular tienen un sello de calidad. Si aumentamos la lista a otros materiales y programas, probablemente el porcentaje de programas con estudios de impacto disponibles sería aún menor.

Pero la historia no termina aquí. La legislación estadounidense ESSA distingue cuatro **niveles de evidencia**. El nivel 1, denominado evidencia fuerte, requiere al menos un resultado favorable y estadísticamente significativo procedente de un estudio experimental bien diseñado y bien ejecutado. En este tipo de estudio, llamado **ensayo controlado aleatorizado** —RCT, por sus siglas en inglés—, la asignación al programa o al grupo de comparación se realiza al azar. El nivel 2, o evidencia moderada, se basa en estudios cuasiexperimentales rigurosos; el nivel 3, o evidencia prometedora, en estudios correlacionales que incorporan controles estadísticos para reducir el sesgo de selección; y el nivel 4 exige, al menos, una justificación teórica acompañada de algún esfuerzo por evaluar sus efectos. La pregunta es, de los estudios disponibles, ¿cuántos han realizado un ensayo aleatorizado alineado con el nivel 1 de ESSA?

Bellwether localizó 101 estudios de esos 17 programas y se quedó con 61 tras aplicar sus criterios. De ellos, aproximadamente uno de cada diez utilizaba un diseño aleatorizado. La mayoría de los estudios restantes utilizaba **diseños cuasiexperimentales**, correlacionales u otros enfoques. Es decir, si miramos a nivel global, el país con más investigación sobre el impacto de programas de matemáticas es Estados Unidos. Dentro de Estados Unidos, la mayoría de esos programas no tienen ningún estudio que evalúe su impacto. Y de los que sí que lo tienen, la gran mayoría utilizan diseños que no permiten afirmar con el mismo grado de confianza que un estudio experimental que el impacto observado se deba a la intervención.

Para complementar estos resultados, realizamos el ejercicio inverso: en lugar de partir de los estudios incluidos en catálogos de evidencia, examinamos qué encuentra una escuela cuando entra en las **páginas de investigación de los propios proveedores**. Con fecha de corte en julio de 2026, recopilamos los documentos que seis proveedores con amplia presencia en el mercado estadounidense —HMH, Curriculum Associates, McGraw Hill, Savvas, Great Minds y Amplify— presentaban en sus portales como investigación o evidencia. En total, codificamos 229 documentos asociados a 31 programas de matemáticas (Innovamat Research Lab, 2026b).

Solo el 7 % de esos documentos correspondía a ensayos aleatorizados. Aproximadamente seis de cada diez no incluían un grupo externo de comparación: presentaban, por ejemplo, resultados antes y después de utilizar el programa, descripciones de su implementación o asociaciones entre el grado de uso y los resultados. Estos análisis correlacionales proporcionan información útil y, dado que organizar un ensayo aleatorizado es muy complejo y caro, es esperable que la mayor parte de lo que se publica sean análisis más sencillos, que también aportan información útil sobre cómo se implementa un programa y cómo se relaciona su uso con los resultados.

Dicho esto, hay que tener presente qué puede y qué no puede responder cada tipo de estudio. Un **análisis correlacional** es más sencillo, pero también menos riguroso: no permite estimar por sí solo qué habría ocurrido con el mismo alumnado si no hubiera utilizado el programa, y sin un **contrafactual** creíble observar una mejora no basta para atribuírsela a la intervención. A esto se añade que, en julio de 2026, 23 de los 31 programas no contaban con una valoración de impacto en *What Works Clearinghouse* o *Evidence for ESSA*. Es decir, ninguna institución externa había revisado sus estudios para clasificarlos en un nivel de evidencia. Nada de esto es un rasgo particular de estas seis editoriales. Describe el estado del sector en general. Generar evidencia causal sobre el propio programa es lento, caro y difícil.

«Generar evidencia causal sobre el propio programa es lento, caro y difícil.»

De 59 programas de matemáticas estadounidenses bien valorados por EdReports, 42 no tenían ni un solo estudio de eficacia publicado en 2023 (Bellwether, 2023). Y en las webs de las seis grandes editoriales, la mayoría de estudios son correlacionales. En otras palabras, son estudios que no pueden detectar el efecto causal de un programa en los resultados académicos de los alumnos.

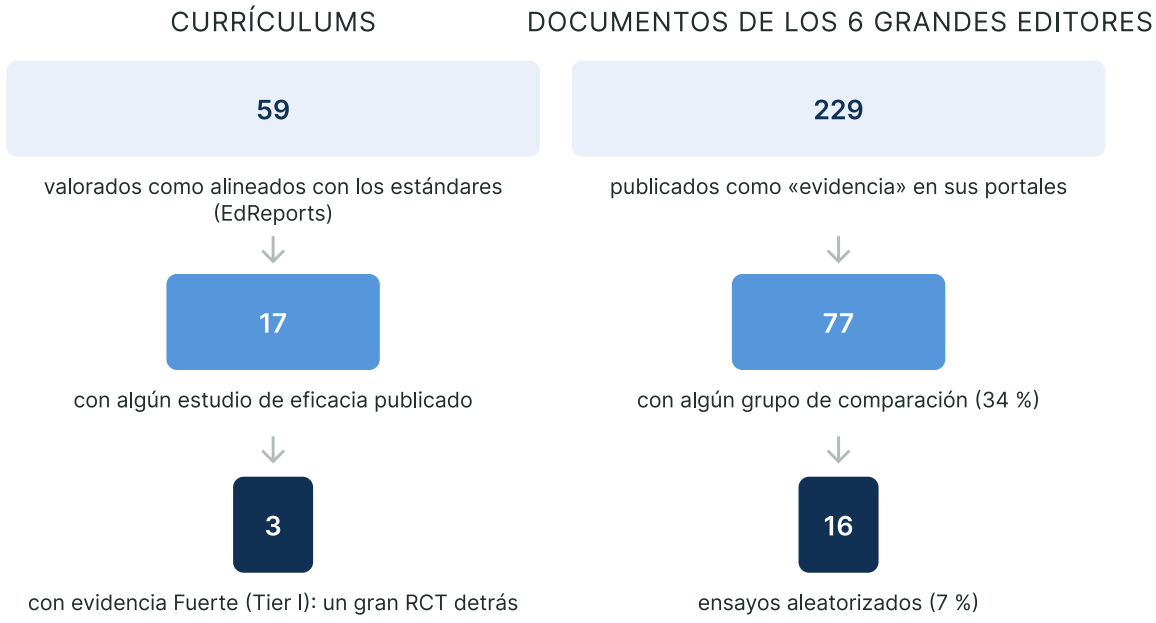


Figura 1. A la izquierda, los programas curriculares evaluados por EdReports, según el estudio de Bellwether (2023). A la derecha, lo que las seis grandes editoriales publican como evidencia, según el estudio de Innovamat Research Lab (2026a, 2026b).

¿Por qué hay tan pocos estudios con un diseño experimental?

No hay una única causa. Entre las explicaciones más habituales se encuentran las siguientes:

- **Coste.** Una evaluación independiente a gran escala puede costar millones: la de *Cognitive Tutor Algebra I* recibió casi 6 millones de dólares del Institute of Education Sciences (2007). En la mayoría de los casos, los estudios experimentales están financiados por fondos públicos.

- **Tiempo.** Un ensayo serio exige varios cursos entre el diseño, la implementación y el análisis; el mismo estudio sobre *Cognitive Tutor Algebra I* no encontró efectos el primer año y sí el segundo (Pane et al., 2014).

- **Acceso a escuelas.** Hace falta convencer a decenas de centros y distritos de que el azar decida qué escuelas o aulas usan el programa. Esto implica que no todos los participantes pueden empezar a utilizarlo desde el inicio, lo que puede generar resistencias y complicaciones logísticas, por ejemplo al tener que gestionar simultáneamente distintas formas de trabajar. Además, pueden surgir dificultades

para obtener la colaboración de las familias y para evitar que los docentes compartan materiales o prácticas entre los grupos de intervención y control, lo que podría reducir las diferencias entre ambos grupos y dificultar la evaluación del impacto real del programa.

- **Coordinación e independencia.** Para aumentar la independencia, es conveniente que alguien externo diseñe, ejecute y publique la evaluación.

- **Incentivos.** Someter un producto ya comercializado a una prueba de resultado incierto tiene pocos alicientes para quien lo vende (Bellwether, 2023).

La evidencia empírica sólida es más la excepción que la norma, aún si nos fijamos en Estados Unidos, el país con más investigación sobre el impacto de programas de matemáticas. La mayoría de programas de matemáticas estadounidenses no tienen ningún estudio que evalúe su impacto. Y de los que sí que lo tienen, la gran mayoría utilizan diseños con limitaciones que no permiten concluir con total confianza que el impacto observado se debe al programa evaluado. Correlación no es causalidad.

2 | Cuando existe evidencia, el efecto suele ser modesto

Cuando un programa de matemáticas se evalúa mediante un diseño capaz de estimar su impacto, ¿qué diferencia suele producir? Cuando la evaluación se focaliza en materiales curriculares concretos, en vez de un programa integral que incluya también formación y acompañamiento, las principales revisiones apuntan a una mejora pequeña, aunque no despreciable.

Slavin y Lake (2008) analizaron trece estudios sobre enfoques curriculares —en su mayoría, libros de texto— y estimaron un efecto medio de +0,10 SD. Años después, Pellegrini et al. (2021) aplicaron criterios de inclusión más estrictos y obtuvieron una estimación menor y no significativa: +0,05 SD en treinta estudios sobre materiales curriculares tradicionales y digitales que incorporaban poca formación docente. Aunque ambas revisiones difieren en el periodo estudiado y en sus criterios metodológicos, ofrecen un orden de magnitud similar. Así pues, la ventaja media de sustituir unos materiales curriculares por otros parece situarse alrededor de +0,05 SD y, en algunas síntesis, cerca de +0,10 SD. Estas cifras comparan los materiales curriculares evaluados con los que ya usaba la escuela, que también enseña matemáticas, y no aíslan la calidad de los materiales: los estudios sobre adopciones reales de libros de texto encuentran pocas diferencias medias entre ellos aunque tengan sellos de calidad distintos (Blazar et al., 2020).

Los estudios a gran escala sobre adopciones reales de libros de texto ofrecen una imagen igualmente matizada. Blazar et al. (2020), tras combinar datos de seis estados estadounidenses después de la implantación de los estándares Common Core, encontraron pocas diferencias medias entre los materiales analizados. Koedel et al. (2017), en cambio, estimaron que uno de los cuatro libros más utilizados en California superaba a los demás en aproximadamente +0,05 a +0,08 SD en 3º de primaria, con resultados también favorables en cursos posteriores. En conjunto, estos estudios sugieren que las diferencias entre materiales curriculares pueden existir, pero no aparecen necesariamente en todos los contextos ni alcanzan siempre la misma magnitud.

La evidencia disponible no es una muestra neutral

Las cifras anteriores describen el efecto medio que aparece en determinadas síntesis de literatura científica, pero la evidencia que llega hasta nosotros no es para nada una muestra neutral de todos los estudios realizados. Algunos resultados tienen más probabilidades que otros de publicarse, incorporarse a un catálogo de evidencia o ser destacados por sus autores.

Este patrón aparece con claridad en nuestro análisis. Codificamos los 85 estudios de las 52 propuestas de matemáticas del catálogo de *Evidence for ESSA* y *What Works Clearinghouse*. Incluyen todo tipo de propuestas, no solo programas de matemáticas: recursos suplementarios (aplicaciones y plataformas digitales), intervenciones focalizadas y tutoría, formación docente y reformas de centro; los programas de matemáticas son 11 de los 52. Ninguno de los 85 estudios presentaba un efecto principal negativo. Entre los 66 que reportan un tamaño del efecto utilizable, la mediana es de +0,15 SD. Por tipo de programa, la mediana es de +0,13 SD tanto en los 15 estudios de programas de matemáticas como en los 27 de aplicaciones y plataformas, de +0,28 SD en las 17 intervenciones focalizadas y tutorías, y de +0,08 SD en los 7 de formación docente y reforma de centro (Innovamat Research Lab, 2026a). Estas cifras son algo superiores a las obtenidas en varias de las síntesis anteriores. Más llamativo aún es que no aparezca ningún resultado principal negativo.

Parte de esta diferencia se explica por la propia forma en que se construyen los catálogos. Los niveles 1, 2 y 3 de ESSA exigen que exista al menos un resultado favorable y estadísticamente significativo y no haya ningún resultado negativo relevante. Por tanto, una muestra formada a partir de programas que alcanzan esos niveles selecciona, de entrada, intervenciones que han obtenido algún resultado positivo. No sería correcto interpretar su efecto medio como si representara al conjunto de programas de matemáticas disponibles.

El contraste con el estudio de Lortie-Forgues e Inglis (2019) ayuda a apreciar esta selección. Los autores analizaron 141 grandes ensayos aleatorizados financiados con fondos públicos en Estados Unidos y

el Reino Unido, con más de 1,2 millones de estudiantes. Las intervenciones evaluadas eran muy diversas y no se habían seleccionado por contar previamente con resultados favorables. Su efecto medio fue de +0,06 SD. Además, el 40 % de los ensayos fue no concluyente: los datos no permitían inclinar claramente la balanza ni hacia la existencia de un efecto ni hacia su ausencia.

A la selección propia de los catálogos se añade el **sesgo de publicación**. Se produce cuando los estudios con resultados positivos o estadísticamente significativos tienen más probabilidades de publicarse que aquellos que encuentran efectos nulos o negativos. En un análisis de 645 estudios procedentes de doce revisiones de investigación educativa, Cheung y Slavin (2016) observaron que los efectos de los artículos publicados eran, en promedio, aproximadamente el doble que los de los documentos no publicados.

La evidencia visible puede estar condicionada, por tanto, por varios filtros relacionados, pero distintos: qué estudios se publican, cuáles se incorporan a los catálogos y qué resultados se seleccionan como principales. Esto no invalida los efectos observados, pero obliga a interpretarlos con prudencia. Los +0,15 SD de nuestra base, o los +0,13 SD de los programas de matemáticas, describen los estudios seleccionados en esos catálogos; no necesariamente el efecto que cabe esperar de cualquier programa de matemáticas.

La misma intervención no siempre tiene el mismo impacto

Como vimos en el capítulo 1, el **tamaño del efecto** no es una etiqueta fija que pueda atribuirse a un programa con independencia de las circunstancias. Además de depender de la alternativa con la que se compara, varía según qué se mide, cuándo se mide y en qué condiciones se implementa. En otras palabras, no existe un único «efecto del programa» en abstracto, sino efectos estimados en contextos y momentos concretos.

El primer factor a tener en cuenta es la **prueba de evaluación** que se utiliza. Algo esencial a tener en cuenta cuando un programa describe su **teoría del cambio** es qué esperamos que pueda cambiar o mejorar, y qué no. En este sentido, la prueba utilizada para evaluar el programa es esencial. Cuando una evaluación mide contenidos, tareas o representaciones muy próximos a los trabajados durante la intervención, los efectos suelen ser mayores que cuando se emplean pruebas más generales, que exigen aplicar lo aprendido a situaciones menos familiares (Cheung y Slavin, 2016). Utilizar pruebas más genéricas permite valorar mejor hasta qué punto el aprendizaje se transfiere más allá de las actividades practicadas.

El estudio de *Fusion* ofrece un buen ejemplo. El efecto fue de g = 0,77 en *ProFusion*, una prueba desarrollada por los investigadores para medir contenidos estrechamente relacionados con la intervención. En exámenes más generalistas, los efectos fueron menores, de entre 0,02 y 0,20 (Clarke et al., 2023). Esta diferencia no significa que no se aprendiera nada, ya que puede reflejar tanto una transferencia todavía limitada como una prueba general poco sensible a lo que la intervención trabajó. Los resultados en la prueba ajustada permiten saber si el alumnado aprendió lo que la intervención pretendía enseñar, y las pruebas independientes informan sobre si ese aprendizaje se traslada a otros tipos de tareas.

En relación con este punto, en nuestro análisis hemos observado una ausencia de medidas que van más allá del aprendizaje (Innovamat Research Lab, 2026b), como por ejemplo las actitudes hacia las matemáticas, la identidad matemática, o la permanencia en la escuela. Los estudios publicados se focalizan en aprendizaje medido a través de pruebas estandarizadas que, en muchos casos, solo capturan una parte de lo que un programa se propone cambiar.

«Los estudios publicados se focalizan en aprendizaje medido a través de pruebas estandarizadas que, en muchos casos, solo capturan una parte de lo que un programa se propone cambiar.»

Un segundo factor a tener en cuenta es el momento en que se evalúan los resultados. Algunas intervenciones producen un impacto positivo justo al terminar que disminuye durante los meses o años posteriores, un fenómeno conocido como ***fade-out***. La investigación indica que esta pérdida del impacto

es frecuente en determinadas intervenciones dirigidas a habilidades cognitivas específicas, especialmente cuando son breves y dejan de aplicarse (Bailey et al., 2017, 2020). Sin embargo, no existe una proporción universal del efecto que se conserve: la evolución depende del tipo de intervención, de los aprendizajes evaluados y de las experiencias posteriores del alumnado.

El programa de tutoría individual *Math Recovery* ilustra este fenómeno. Al finalizar 1º de primaria, produjo efectos de entre +0,15 y +0,30 SD en pruebas administradas por evaluadores externos, y efectos todavía mayores en una prueba estrechamente vinculada al programa. Al final de 2º, en cambio, las estimaciones eran pequeñas y ninguna podía distinguirse estadísticamente de cero (Smith et al., 2013). Esto no demuestra que no quedara ningún aprendizaje, sino que, un curso después, el estudio ya no podía establecer una ventaja fiable del grupo que había recibido la intervención frente al grupo de comparación.

El tercer factor son las **condiciones de implementación**. Podemos imaginar los materiales curriculares como una partitura y al profesorado como quien la interpreta. Sin intérprete no hay música, y un mismo intérprete no obtiene lo mismo de cualquier partitura. Unos materiales pueden facilitar tareas más ricas, trayectorias de aprendizaje más coherentes, mejores conexiones entre ideas y más oportunidades de práctica y retroalimentación. Otros pueden dificultar todo ello. Aun así, el resultado también depende de cómo comprende el profesorado la propuesta, cuánto tiempo puede dedicarle, cómo responde a las necesidades del alumnado y qué formación y acompañamiento recibe para llevarla al aula.

Por tanto, el efecto observado en un estudio no es únicamente el resultado de los materiales. Es el resultado conjunto del programa, de las personas que lo utilizan y de las condiciones organizativas que hacen posible su implementación. Dos escuelas pueden adoptar los mismos materiales curriculares y obtener resultados distintos si difieren en el tiempo disponible, el apoyo al profesorado, la estabilidad de los equipos o el grado en que consiguen incorporar la propuesta a la práctica cotidiana. Sobre estos aspectos hablamos más en profundidad en el artículo «¿De qué depende aprender matemáticas?».

«Dos escuelas pueden adoptar los mismos materiales curriculares y obtener resultados distintos si difieren en el tiempo disponible, el apoyo al profesorado, la estabilidad de los equipos o el grado en que consiguen incorporar la propuesta a la práctica cotidiana.»

En resumen, el impacto de una misma intervención depende de cuatro cosas: con quién se compara (cuanto más parecido sea el grupo de comparación y más similares sean las actividades que realiza, más difícil será detectar un efecto positivo), con qué prueba se evalúa (una prueba propia y alineada con el programa dará un efecto mayor que una externa y general), cuándo se evalúa (cuanto más tiempo pasa desde la intervención, menor suele ser el efecto medido) y cómo se ha implementado.

Esta cuestión cobra especial importancia cuando la intervención se aplica a gran escala. Un programa puede obtener buenos resultados en condiciones especialmente favorables —con centros seleccionados, un seguimiento estrecho y un apoyo intensivo de sus creadores— y no reproducirlos con la misma magnitud cuando se extiende a contextos más diversos y funciona con los recursos habituales del sistema. Por eso, las directrices del Institute of Education Sciences y la National Science Foundation distinguen entre los **estudios de eficacia** y los de **efectividad**. Los primeros analizan si una intervención puede funcionar en condiciones favorables y relativamente controladas. Los segundos estudian si continúa produciendo resultados en condiciones ordinarias, con poblaciones diversas y sin una participación de los creadores mayor que la que cabría esperar en una implementación habitual (IES y NSF, 2013).

Apuntes finales

La idea que repite esta sección es que la misma intervención no siempre tiene el mismo impacto: la medida de efecto no se tiene que interpretar como una propiedad fija de un programa, sino en relación con la pregunta que el estudio pretende responder: ¿Qué diferencia supone utilizar este programa en lugar de la alternativa disponible, para un resultado concreto, una determinada población y un contexto específico? Por eso, es importante recordar algo que comparten buena parte de los estudios sobre

materiales curriculares. El grupo de comparación no deja de recibir enseñanza, sino que continúa utilizando los materiales habituales. Por tanto, el tamaño del efecto no representa todo lo que aprende el alumnado con el nuevo programa, sino su ventaja adicional respecto a otra propuesta con la que también estaría aprendiendo matemáticas. Esta comparación exigente ayuda a entender por qué los tamaños del efecto en estos estudios se sitúan entre el +0,05 SD y +0,10 SD. Según los puntos de referencia propuestos por Kraft (2020, 2023), +0,10 SD se sitúa en el tramo intermedio de los efectos observados en estudios causales de intervenciones educativas medidos mediante pruebas estandarizadas.

Y aquí viene la pregunta clave: ¿es suficiente esta mejora? ¿Vale la pena la inversión? Como argumenta Whitehurst (2009) tras comparar el coste y el beneficio de múltiples intervenciones educativas, cambiar los materiales curriculares puede ser una palanca eficiente, porque los programas más eficaces no tienen necesariamente un precio de adquisición superior al de las alternativas. Modesto no significa irrelevante. Escoger un material curricular u otro apenas cambia el coste, pero puede producir una mejora moderada que llega a todo el grupo.

«Modesto no significa irrelevante. Escoger un material curricular u otro apenas cambia el coste, pero puede producir una mejora moderada que llega a todo el grupo.»

En el mejor de los casos, el efecto típico de cambiar de materiales curriculares está entre +0,05 y +0,10 SD, aunque este efecto puede disminuir con el tiempo o cuando la intervención se implementa a gran escala. Aun así, cambiar los materiales curriculares puede ser una palanca efectiva, debido a que el coste de programas alternativos es, a menudo, muy parecido.

3 | Los materiales curriculares rinden cuando no viajan solos

Los materiales curriculares rinden más cuando llegan con formación

Cambiar los materiales en el aula de matemáticas produce, por sí solo, un efecto pequeño. Como ya hemos visto, el meta-análisis de Pellegrini et al. (2021) sitúa el impacto de los materiales curriculares que llegan sin apenas formación en +0,05 SD, un efecto no significativo. Cuando esos mismos materiales van acompañados de formación y apoyo a la implementación, el efecto sube a +0,12 SD. El meta-análisis de Lynch et al. (2019), sobre 95 estudios de matemáticas y ciencias, apunta en la misma dirección: los programas que incorporan solo materiales o solo formación obtienen alrededor de +0,15 SD, mientras que los que combinan ambas cosas alcanzan +0,25 SD. Los materiales por sí solos cambian poco el aprendizaje. La mejora aparece cuando llegan acompañados de formación específica para el profesorado.

Las dos cifras no son intercambiables. La de Pellegrini (+0,12 SD) proviene de siete estudios de primaria, solo de matemáticas, en los que la formación acompañaba a un libro de texto tradicional, y se compara con los +0,04 SD de esos mismos materiales con poca formación. La de Lynch (+0,25 SD) mezcla matemáticas y ciencias, todas las etapas y programas que combinan formación y materiales de cualquier tipo, e incluye pruebas diseñadas por los propios investigadores, que dan efectos unos 0,27 SD mayores que las estandarizadas. Lo que ambas comparten es la diferencia: la **formación ligada a los materiales curriculares** añade entre +0,08 y +0,10 SD respecto a los materiales solos.

Merece la pena precisar qué entienden estos meta-análisis por formación. En Pellegrini et al. (2021), los materiales curriculares «con poca formación» son aquellos en los que el profesorado recibe menos de dos días o 15 horas de formación, poco más que una introducción a los materiales; la categoría que alcanza +0,12 SD agrupa programas con al menos dos días o 15 horas de formación dedicada a implementar unos materiales curriculares concretos en el aula, seguida de acompañamiento (*coaching*) durante el curso. En Lynch et al. (2019), la formación incluye talleres de verano, sesiones de seguimiento durante el curso, *coaching* y trabajo entre docentes del mismo centro, y los materiales curriculares son programas que introducen tareas y secuencias nuevas con orientaciones para el docente incorporadas en el propio material. Un ejemplo de programa que combina ambas cosas es una unidad de fracciones acompañada de un taller de cinco días en verano y trece sesiones en la escuela durante el curso (Saxe et al., 2001); otro, *SimCalc*, una unidad de proporcionalidad para secundaria con software y un taller de formación para el profesorado, evaluada en tres ensayos a gran escala en Texas (Roschelle et al., 2010). En el análisis de Lynch, la formación rinde más cuando se centra en cómo usar los materiales y en cómo aprende el alumnado el contenido, cuando incluye reuniones de seguimiento de la implementación y cuando participan docentes del mismo centro. El número de horas, en cambio, no predice el efecto, y los componentes en línea se asocian a efectos menores.

Uno podría pensar que la formación docente es lo que realmente crea el impacto. Pero tres grandes ensayos aleatorizados contradicen, en parte, esta intuición: una formación intensiva en contenido matemático, pero desligada del aula, consiguió que los profesores supieran más matemáticas sin que se viera ningún efecto positivo detectable en sus alumnos (Garet et al., 2016); en los dos ensayos anteriores de la misma serie no mejoró ni siquiera el conocimiento docente (Garet et al., 2010, 2011). El efecto de la formación parece llegar al alumnado cuando cambian las prácticas reales que se realizan en el aula de matemáticas (Lynch et al., 2025). Aquella formación que hace pensar al docente sobre su propia aula y le da recursos que puede llevar a la práctica, tiene un mayor impacto que las que se limitan a prescribir conductas o añadir contenido sin ejemplos reales que el docente pueda adaptar a su aula (Kennedy, 2016).

«El efecto de la formación parece llegar al alumnado cuando cambian las prácticas reales que se realizan en el aula de matemáticas»

Cómo se compara con otras palancas: efecto, coste y escala

Hasta aquí hemos comparado unos mismos materiales curriculares con y sin formación. La segunda pregunta es distinta: ¿cómo se compara esa combinación con otras formas de mejorar el aprendizaje que una escuela puede contratar, como el *coaching* o la tutoría, en efecto, en coste y en escala?

La primera alternativa es el ***coaching***, el acompañamiento individual en el aula de un experto que observa y prepara clases con el docente durante meses (como explicamos más en detalle en el capítulo 3). Es la forma de formación docente con más efecto conocido sobre la práctica en el aula: +0,49 SD (Kraft et al., 2018). Su efecto sobre el aprendizaje del alumnado es menor, +0,18 SD, y procede casi por completo de intervenciones de lectura; en matemáticas la evidencia es todavía escasa y no significativa. Además, no hay que verlo como una palanca separada de los materiales: la formación ligada a los materiales curriculares que alcanza +0,12 SD en primaria incluye ese acompañamiento como componente. Volviendo a nuestra partitura: la formación general es un cursillo de dirección de orquesta; el *coaching* es ensayar semana a semana con un maestro al lado del atril.

Construir capacidad instruccional optimiza la eficacia y la escalabilidad

Mejores materiales y mejor apoyo docente se potencian mutuamente. Cuanto más se construye el sistema para empoderar al profesor, mayores son los beneficios.

* Por claridad visual, cada intervención se sitúa en el mayor tamaño del efecto observado en la literatura revisada; las estimaciones más bajas de los mismos estudios aparecen en la columna «Tamaño medio del efecto».

Figura 2. Comparación del tamaño del efecto y la escalabilidad de distintas intervenciones educativas, ordenadas de menor a mayor efecto estimado. Nota: las plataformas digitales de práctica quedan excluidas del gráfico por falta de comparabilidad: la mayoría de los tamaños del efecto reportados tienen en cuenta a los estudiantes que las utilizan con fidelidad de implementación, y no a todos los estudiantes del aula. En cuanto al coaching, la mayoría de los programas incluidos en Kraft et al. (2018) se centran en el acompañamiento docente en lectura, no en matemáticas.

La segunda alternativa, y la de mayor efecto directo sobre el alumnado, es la **tutoría** en grupos muy pequeños y dentro del horario escolar, con un efecto medio de +0,29 SD para el conjunto de programas de tutoría en un meta-análisis de casi un centenar de ensayos aleatorizados (Nickow et al., 2024), y de +0,40 SD en la síntesis más amplia, de 263 ensayos (Kraft et al., 2026). Es la intervención más potente que conoce la investigación educativa. Tanto la tutoría como el *coaching* tienen algo en común. Tienen un coste muy elevado, porque requieren una intervención muy individualizada, con muchos recursos humanos y mucha implicación sostenida en el tiempo. Esto hace que la solución sea difícilmente escalable a muchos contextos. Cuanto más se acerca una intervención al alumno concreto, más cuesta y peor escala (Figura 2). La tutoría exige una persona por cada tres o cuatro alumnos. El *coaching* exige un experto por cada pocos docentes, y sus efectos caen a la mitad o más cuando los programas crecen: de +0,28 SD en pilotos a +0,10 a gran escala (Kraft et al., 2018). La tutoría, medida en implantaciones grandes con tests externos, baja a +0,16–0,22 (Kraft et al., 2026). Los materiales curriculares recorren el arco en sentido contrario: su efecto aislado es el más pequeño, pero llega a todas las aulas a la vez, cada día, a coste casi nulo. Y la formación ligada a los materiales, la que enseña a usar los materiales curriculares concretos que el docente tiene en el aula, ocupa el punto de equilibrio del arco: aumenta el efecto de los materiales sin renunciar a su escala. Solo la tutoría supera claramente en efecto a esa combinación; el *coaching* la iguala, a un coste por docente mucho mayor.

«Cuanto más se acerca una intervención al alumno concreto, más [recursos] cuesta y peor escala.»

El núcleo pedagógico

Paremos un momento a pensar qué tienen en común todos estos resultados. La formación que funciona es la que cambia el aula. La tutoría que funciona es la que se liga a los materiales curriculares de cada día. Los materiales que funcionan son los que alguien sabe usar. Todo apunta al mismo sitio: el aprendizaje no se produce en el material, ni en el docente, ni en el alumno por separado, sino en la interacción entre los tres. Es lo que la literatura define como **«núcleo pedagógico»** (Figura 3).

El aprendizaje se produce en el «núcleo pedagógico»: la interacción entre docentes y alumnos alrededor de los materiales. Según esta interacción, mejorar un solo factor mientras se ignoran los otros dos rinde poco: el impacto es mayor cuando la intervención o programa los tiene en cuenta todos.

Figura 3. El núcleo pedagógico funciona como una interacción con tres factores: si un factor falla, el conjunto rinde poco. A partir de Cohen et al. (2003).

La misma lógica se repite fuera de Estados Unidos. El panel de expertos del Banco Mundial señala la **«pedagogía estructurada» —**materiales de calidad más formación y acompañamiento docente, todo junto **—** como una de las inversiones educativas más rentables que existen (GEEAP, 2023). No es un programa concreto: es exactamente la combinación de la que habla esta sección.

Hagamos un resumen de lo aprendido en esta sección. Primero, cambiar de materiales curriculares puede producir efectos pequeños o moderados. Segundo, si estos cambios van acompañados de formación ligada a la práctica, los tamaños del efecto crecen. Tercero, hay una tensión entre **impacto y escala**: la tutoría y las intervenciones en pequeño grupo, las palancas con más efecto, y el *coaching*, la formación más intensiva, son las que menos escalan, así que su papel es complementar, no sustituir.

Pero aquí la analogía de la orquesta llega a su límite: en una escuela, el «público» también toca. Los alumnos no reciben el programa: lo hacen, con su punto de partida, su familia y su contexto. Y alrededor del aula hay más músicos de los que parece: la dirección del centro que decide horarios y prioridades, las políticas educativas que fijan el currículo oficial y los recursos, los compañeros, las familias. Este capítulo ha mirado la partitura y sus ensayos; el capítulo 3 abre el plano entero: qué factores determinan el aprendizaje de las matemáticas, cuáles son modificables y cuáles vienen dados, y qué puede hacer cada actor para que la obra suene lo mejor posible (capítulo 3 «¿De qué depende aprender matemáticas?»).

Los materiales curriculares rinden entre +0,05 y +0,10 SD; con formación docente ligada a la práctica, entre +0,12 y +0,25 SD según la síntesis. La tutoría y las intervenciones en pequeño grupo tienen efectos mayores, y el coaching mejora sobre todo la práctica docente, pero ambos cuestan más y escalan peor: actúan como complemento, no como solución única. El equilibrio entre efecto y escala está en los programas que combinan buenos materiales con formación ligada a la práctica.

4 | Qué implica esto en la práctica

1. Ajustar las expectativas. Un efecto sostenido de 0,10–0,20 SD es de lo mejor que ofrece este campo: el alumnado llega al mismo sitio unas semanas o unos meses antes. Hablamos de mejoras relevantes pero graduales, no de transformaciones radicales en un solo curso. Los avances educativos suelen ser modestos, acumulativos y dependientes del contexto, por lo que las promesas de cambios espectaculares deben analizarse con cautela.

2. Pedir estudios y hacer preguntas. Los estudios de impacto y las evidencias nos ayudan a tomar decisiones más informadas. Es importante pedir evidencias a los desarrolladores y analizarlas con detalle. ¿Cuál es el tamaño del efecto y cómo se compara con la referencia de 0,10 SD? Si los estudios no existen o son poco transparentes con el diseño y la medida del efecto, pregúntate por qué. Además, es importante ser críticos cuando leemos estudios, como vimos en el capítulo 1: ¿Quién firma el estudio, el desarrollador o un evaluador externo? ¿Qué diseño se ha utilizado y cómo se ha estimado que hubiera pasado sin el programa? ¿Con qué prueba se ha evaluado el programa?

3. Valorar el paquete completo de un programa, no solo los materiales. La formación y el acompañamiento que viajan con el material pueden pesar más que el material mismo. En el meta-análisis de Lynch et al. (2019), los programas que incorporan solo materiales o solo formación obtienen alrededor de +0,15 SD, y los que combinan ambas cosas alcanzan +0,25 SD. En los estudios de primaria de Pellegrini et al. (2021) el contraste es equivalente: +0,05 SD sin apenas formación y +0,12 SD con apoyo a la implementación. Las dos síntesis miden cosas distintas (primaria y pruebas externas en un caso; matemáticas y ciencias, con pruebas de los propios investigadores, en el otro), pero coinciden en lo esencial: la formación ligada a los materiales curriculares añade entre +0,08 y +0,10 SD. Preguntar qué tipo de formación incluye debería ser tan normal como hojear las guías para los docentes o los libros para el alumnado.

5 | Preguntas frecuentes

«Pocos programas tienen evidencia fuerte. ¿Cuáles son los que sí?»

Tres programas de matemáticas tienen hoy la certificación más alta (evidencia fuerte) del catálogo *Evidence for ESSA*. *Math Expressions*: +0,11 SD en el mayor ensayo curricular público de EE. UU., más de un centenar de escuelas asignadas al azar y medidas con un test externo (Agodini et al., 2010, 2013). *Math in Focus*: ≈+0,12 SD en un ensayo aleatorizado en Nevada, encargado por la propia editorial a una consultora externa, con efecto nulo en el test estatal (Jaciw et al., 2016). Y el álgebra de *Carnegie Learning*: efecto nulo el primer año y de aproximadamente +0,20 SD en el segundo, en el ensayo de RAND en siete estados (Pane et al., 2014).

«El programa que estamos mirando no aparece en What Works Clearinghouse. ¿Mala señal?»

No necesariamente. Para empezar, los programas no estadounidenses rara vez están listados, simplemente porque para obtener la certificación uno tiene que presentar estudios desarrollados dentro de Estados Unidos. En este caso, es necesario mirar si tienen algún estudio publicado en una revista científica y analizar cómo han definido el grupo comparación, con qué prueba se mide el estudio, cuando y en qué contexto. Por otra parte, los dos grandes registros estadounidenses (*What Works Clearinghouse* y *Evidence for ESSA*) se solapan mucho menos de lo que parece: de los 52 programas del catálogo *Evidence for ESSA*, solo 3 tienen además informe del *What Works Clearinghouse*, y a la inversa hay programas con buena evidencia que no están en el catálogo —*Building Blocks*, en Educación Infantil, tiene ensayos aleatorizados revisados por pares y no está certificado—. También hay ausencias elocuentes: el WWC revisó *enVision*, uno de los programas más vendidos del país, y no encontró ni un solo estudio que cumpliera sus estándares.

«¿Habéis analizado las evidencias de las grandes editoriales estadounidenses? ¿Qué habéis encontrado?»

Sí, hemos analizado en profundidad las evidencias públicas de seis grandes editoriales. Recopilamos y codificamos uno a uno los 229 documentos que seis grandes editoriales presentan como investigación o evidencia en sus propios portales (en julio de 2026). Describiremos el tipo de evidencia que cada editorial decide publicar en su web, no la calidad de sus programas: una editorial puede tener un buen programa y comunicar poco sobre su impacto, y publicar mucho material no garantiza que ese material permita estimar un efecto.

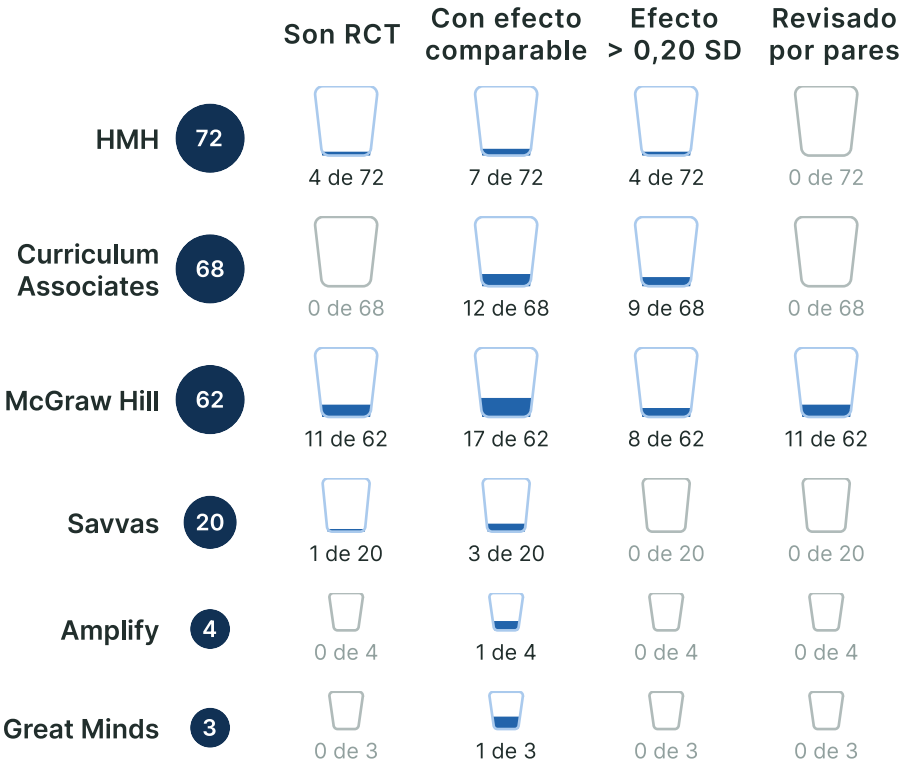
HMH es la que más publica, con 72 documentos en su portal de investigación, más de la mitad de ellos comparaciones antes y después de utilizar el programa, sin grupo de comparación. Curriculum Associates publica 68 documentos, en torno al 60 % análisis correlacionales, y ninguno corresponde a un ensayo aleatorizado. McGraw Hill publica 62 y es la única del grupo con artículos revisados por pares, concentrados en dos programas: los ensayos aleatorizados de *Building Blocks* encontraron efectos positivos y sustanciales, y están entre los estudios curriculares más sólidos publicados en revistas académicas, mientras que los tres de *ALEKS* no encontraron efectos estadísticamente significativos. Savvas publica 20 documentos, casi todos con grupo de comparación, aunque solo 3 incluyen un tamaño del efecto, que es precisamente lo que nos permite valorar la magnitud del impacto y compararlo con el de otras opciones. Great Minds publica 3 documentos sobre *Eureka Math* y Amplify, 4. Recordemos que un volumen menor no implica por sí mismo una base de evidencia más débil, ya que puede reflejar una política más selectiva sobre qué se publica. En conjunto, el 7 % de los documentos corresponde a ensayos aleatorizados, el 18 % reporta un tamaño del efecto estandarizado frente a un grupo de comparación y el 5 % ha pasado por revisión por pares. La base de datos completa, con los resultados de cada estudio, está disponible de forma abierta (Innovamat Research Lab, 2026b) (Figura 4).

Vale la pena poner estas cifras en contexto antes de sacar conclusiones. Un ensayo aleatorizado en condiciones reales es un ejercicio costoso y lento: exige convencer a decenas de centros para que acepten que el azar decida qué programa utilizarán, sostener el estudio durante varios cursos, financiar una evaluación independiente y aceptar de antemano que el resultado puede no ser favorable. La evaluación de *Cognitive Tutor Algebra I* que realizó RAND en siete estados costó cerca de seis millones de dólares y se prolongó durante años (Institute of Education Sciences, 2007; Pane et al., 2014). Ninguna editorial puede poner en marcha un estudio así por su cuenta, y aunque quisiera el coste de hacerlo es a menudo astronómico. A esto se añade que nuestro recuento describe lo que cada editorial publica en su

web en una fecha concreta, no todo lo que existe: puede haber estudios en curso, investigaciones no publicadas o evaluaciones realizadas por terceros que no aparecen en esos portales. Además, una vez realizado el estudio, publicarlo en revistas científicas revisadas por pares aún aumenta más la lentitud y complejidad del proceso y, de hecho, muchos estudios de impacto válidos nunca llegan a una revista académica.

Las grandes editoriales: muchos estudios, muy poca evidencia comparable

229 documentos localizados (agosto de 2026); solo 41 miden el impacto con un tamaño de efecto estandarizado frente a un grupo de comparación, y solo 11 han pasado revisión por pares. Cada cubo se llena según la proporción.



«Efecto comparable» = tamaño de efecto estandarizado (d o g) procedente de un diseño con grupo de comparación; se excluyen los pre/post sin control, los coeficientes beta y los puntos porcentuales. En Curriculum Associates, la mayoría de los efectos > 0,20 SD se miden con el test de la propia editorial (i-Ready Diagnostic).

Fuente: Innovamat Research Lab (2026b); 229 documentos de los portales de investigación de las seis grandes editoriales de EE. UU., recopilados en agosto de 2026.

Figura 4. Documentos sobre 31 programas de matemáticas publicados en los portales de investigación de seis grandes editoriales (229 documentos, julio de 2026). Las editoriales publican muchos documentos y muy poca evidencia comparable: solo 41 de los 229 reportan un tamaño de efecto estandarizado frente a un grupo de comparación, y solo 11 pasan revisión por pares. En Curriculum Associates, la mayoría de los efectos grandes se miden con el test de la propia editorial. Elaboración propia (Innovamat Research Lab, 2026b).

«¿No es mejor un estudio de la propia editorial que ningún estudio?»

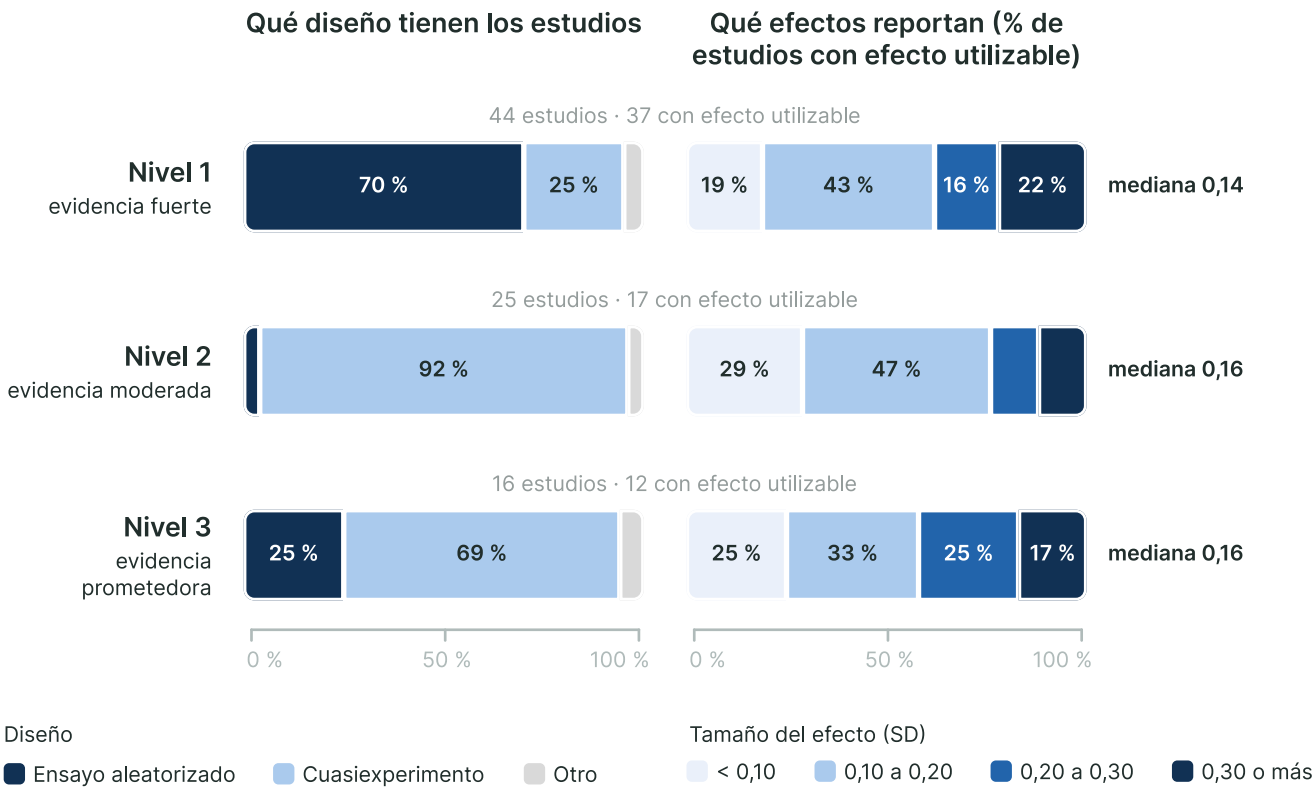
Sí, siempre que se lea sabiendo lo que es. Los datos de algunas investigaciones dicen dos cosas sobre los estudios internos. Primera: reportan menos y peor —casi la mitad no ofrece ningún tamaño de efecto, solo afirmaciones de mejora—. Segunda: tienden a medir con una prueba propia, ajustada al contenido trabajado, y eso produce cifras sistemáticamente mayores que las de una prueba externa. Los propios estudios lo muestran con claridad, porque aplican las dos medidas al mismo alumnado y en el mismo periodo: *Math 180* reporta 0,51 SD en el test de su editorial y 0,32 en el externo; *Math Recovery*, 1,04 en su medida y 0,28 fuera; *ROOTS*, 0,81 frente a 0,23; *Fusion*, 0,77 frente a 0,07, este último no significativo. Ahora bien, una diferencia así no demuestra que no se aprendiera nada: puede que el aprendizaje aún no se transfiera a tareas menos familiares, o que la prueba general, mucho más amplia, no sea lo bastante sensible para detectarlo. Por eso hay que saber siempre qué prueba se está mirando.

«¿Cómo se reparten los diseños y los efectos dentro del catálogo certificado?»

Hay una variación grande, sobre todo entre niveles de evidencia. Nuestra base de datos cubre 52 programas curriculares de todo tipo y 85 estudios del catálogo de matemáticas de *Evidence for ESSA*. El 44 % de los estudios son ensayos aleatorizados, el 53 % cuasiexperimentos y un 3 % otros diseños, y el nivel de certificación explica casi a la perfección el tipo de diseño: entre los estudios de programas de nivel 1, el 70 % son aleatorizados (no el 100 %, porque la etiqueta es del programa, no del estudio: basta un buen RCT y los demás estudios del mismo programa quedan amparados bajo ella), mientras que en el nivel 2 el 92 % son cuasiexperimentos. De los 85 estudios, 19 no reportan ningún tamaño de efecto utilizable; entre los 66 restantes, la mediana es +0,15 SD y los diseños más débiles reportan medianas más altas (≈0,16 en el nivel 3 frente a ≈0,14 en el nivel 1): menos rigor, cifras mayores. Las apps, el grupo más numeroso del catálogo (20 de 52), se agrupan en +0,07–0,13 SD; los efectos grandes se concentran en las intervenciones de pequeño grupo y la tutoría, y un tercio de los efectos de 0,25 SD o más se debe a pruebas alineadas con el programa (Figura 5).

Menos rigor, cifras mayores: el catálogo de Evidence for ESSA por nivel de evidencia

85 estudios y 52 programas de matemáticas · el tramo de cada efecto se calcula sobre los estudios que reportan un tamaño del efecto utilizable



Fuente: análisis propio del catálogo de matemáticas de Evidence for ESSA (Innovamat Research Lab, 2026a). La etiqueta de nivel es del programa, no del estudio: por eso hay cuasiexperimentos en el nivel 1.

Figura 5. Diseño de los estudios y distribución de los tamaños del efecto del catálogo de matemáticas de Evidence for ESSA, por nivel de evidencia. Incluye todos los tipos de programa del catálogo (programas de matemáticas, apps y plataformas, intervenciones y tutoría, formación docente). Los porcentajes de la derecha se calculan sobre los 66 estudios con un tamaño del efecto utilizable. Elaboración propia (Innovamat Research Lab, 2026a).

«¿Y las apps y plataformas digitales de calidad? ¿Son una buena inversión?»

Los datos apuntan a que sí, pero con matices. Sus efectos (+0,07–0,13 SD) son parecidos a los de un cambio de materiales curriculares y su coste suele ser mucho menor, así que, bien elegidas, pueden ser inversiones eficientes. La letra pequeña: Primero, muchos estudios de aplicaciones solo cuentan al alumnado que las usó «con fidelidad» —los usuarios asiduos—, no a todos los que las recibieron. Eso tiende a inflar la cifra, porque quien usa mucho una herramienta no es un alumno cualquiera, y los análisis

6 | Para llevarse

Tres frases resumen el capítulo. Primero, la evidencia es la excepción. La mayoría de los programas de matemáticas no tienen ningún estudio y muy pocos tienen estudios capaces de demostrar con confianza alta que cualquier cambio observado se debió al uso del programa. Segundo, los efectos reales son modestos (en torno a una décima de desviación estándar), pero a menudo pueden convertirse en una buena inversión, si el coste de los productos alternativos es similar. Finalmente, la palanca no son solo los materiales curriculares, sino el paquete que puede ofrecer un buen programa: su efecto casi se duplica con creces cuando llegan con formación ligada a la práctica. El capítulo 3 explica el porqué de fondo: recorre los factores que determinan el aprendizaje —de la política educativa a la familia, con el docente en el centro— y señala cuáles puede mover una escuela.

Analiza la evidencia de aquello que quieres implementar, espera mejoras modestas y prioriza aquellos materiales curriculares con un buen apoyo a la implementación.

Para profundizar

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

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). Achievement effects of four early elementary school math curricula: Findings for first and second graders (NCEE 2011-4001). Institute of Education Sciences.

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). After two years, three elementary math curricula outperform a fourth (NCEE 2013-4019). Institute of Education Sciences.

Bailey, D. H., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10(1), 7–39.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & 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.

▸ Bellwether Education Partners (2023). Rounding up: An analysis of math curriculum effectiveness studies. Bellwether.

Blazar, D., Heller, B., Kane, T., Polikoff, M., Staiger, D., et al. (2020). Curriculum reform in the common core era: Evaluating elementary math textbooks across six US states. Journal of Policy Analysis and Management, 39(4), 966–1019.

Carlana, M., & La Ferrara, E. (2025). Apart but connected: Online tutoring, cognitive outcomes, and soft skills. American Economic Review, 115(10), 3487–3513.

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

Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. Journal of Research on Educational Effectiveness, 16(2), 326–349.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. Educational Evaluation and Policy Analysis, 25(2), 119–142.

Garet, M. S., et al. (2010, 2011). Middle school mathematics professional development impact study (NCEE 2010-4009; NCEE 2011-4024). Institute of Education Sciences.

Garet, M. S., et al. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010). Institute of Education Sciences.

GEEAP — Global Education Evidence Advisory Panel (2020, 2023). Cost-effective approaches to improve global learning. Banco Mundial.

Gortazar, L., Hupkau, C., & Roldán-Monés, A. (2024). Online tutoring works: Experimental evidence from a program with vulnerable children. Journal of Public Economics, 232, 105082.

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

Innovamat Research Lab (2026a). Math program efficacy studies: base de datos abierta de evidencia de eficacia de programas de matemáticas en EE. UU. https://docs.google.com/spreadsheets/d/1c4LJ_Kyt0oQt-inIFzhyEdQ8rkiPOn1jNEfApO2wHB0

Innovamat Research Lab (2026b). Publisher efficacy claims: corpus de los documentos de eficacia publicados en los portales de las seis grandes editoriales de matemáticas de EE. UU. https://docs.google.com/spreadsheets/d/1XACeSy6m91OFSLLSx-IzNNdkmIlGhTTe196IroAMI3M

Institute of Education Sciences (2007). Effectiveness of Cognitive Tutor Algebra One implemented at scale (Award R305A070185). U.S. Department of Education.

Institute of Education Sciences & National Science Foundation (2013). Common guidelines for education research and development. U.S. Department of Education y National Science Foundation.

Jaciw, A. P., Hegseth, W. M., Lin, L., Toby, M., Newman, D., Ma, B., & Zacamy, J. (2016). Assessing impacts of Math in Focus, a "Singapore Math" program. Journal of Research on Educational Effectiveness, 9(4), 473–502.

Jerrim, J., & Vignoles, A. (2016). The link between East Asian "mastery" teaching methods and English children's mathematics skills. Economics of Education Review, 50, 29–44.

Kennedy, M. M. (2016). How does professional development improve teaching? Review of Educational Research, 86(4), 945–980.

Koedel, C., Li, D., Polikoff, M. S., Hardaway, T., & Wrabel, S. L. (2017). Mathematics curriculum effects on student achievement in California. AERA Open, 3(1), 2332858417690511.

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., & 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.

Kraft, M. A., Schueler, B. E., & Falken, G. T. (2026). What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability. Review of Educational Research. Publicación anticipada en línea.

Lortie-Forgues, H., & 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., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. Educational Evaluation and Policy Analysis, 41(3), 260–293.

Lynch, K., Gonzalez, K. E., Hill, H. C., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction, and student achievement. AERA Open, 11, artículo 23328584251335302.

Nickow, A. J., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for preK–12 learning: A systematic review and meta-analysis of the experimental evidence. American Educational Research Journal, 61(1), 74–107.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. Educational Evaluation and Policy Analysis, 36(2), 127–144.

Pane, J. F., Doss, C. J., Todd, I., & Seaman, D. (2025). Efficacy of Zearn Math over two years in grades 3 to 5: An experiment in Texas (EdWorkingPaper 25-1211). Annenberg Institute, Brown University.

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

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. American Educational Research Journal, 47(4), 833–878.

Saxe, G. B., Gearhart, M., & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. Journal of Mathematics Teacher Education, 4(1), 55–79.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. Review of Educational Research, 78(3), 427–515.

que cuentan a todos los alumnos asignados encuentran efectos mucho menores. Segundo, la evidencia más reciente apunta a la parte baja del rango: el mayor ensayo aleatorizado reciente de una plataforma (*Zearn*, en Texas) encontró +0,07 SD —no significativo— en el test estatal y +0,11 en un test intermedio (Pane et al., 2025). La regla del capítulo aplica también aquí: pedir el efecto sobre todos los alumnos asignados, medido con un test externo.

«¿Quién hace todos estos estudios?»

Mayoritariamente universidades y consultoras de investigación. En el catálogo certificado, un 35 % lo firman equipos universitarios. En particular, el laboratorio de Fuchs en Vanderbilt y el grupo de Clarke en Oregón concentran los ensayos de intervención, mientras que el CRRE de Johns Hopkins trabaja por encargo de las editoriales. Aparte, un 40 % lo firman consultoras de investigación (LXD, Mathematica, SRI, AIR, RMC, Empirical Education, WestEd, RAND, HumRRO, etc.) y un 20 % los equipos internos de las editoriales. Solo un tercio ha pasado revisión por pares. En las webs de las editoriales la proporción se invierte: dos de cada tres documentos los firma la propia casa. Y un dato estructural a destacar: casi todos los ensayos aleatorizados independientes de programas de matemáticas existen gracias a financiación pública (sobre todo, con financiación del Institute of Education Sciences).

«¿Esto pasa igual fuera de Estados Unidos?»

La escasez es aún mayor. Fuera de EE. UU. casi no hay evidencia sobre programas concretos, y la investigación se centra en enfoques pedagógicos. Pero las conclusiones no cambian demasiado. En el Reino Unido, el programa *Mathematics Mastery* produjo +0,07 SD al combinar los dos ensayos de la fundación EEF (Jerrim y Vignoles, 2016) —la cifra típica de un cambio de materiales curriculares—. En países de renta baja y media, la «pedagogía estructurada» (materiales más formación y acompañamiento) es una de las apuestas más rentables del panel del Banco Mundial (GEEAP, 2023). Y dos ensayos europeos recientes de tutoría en línea —en torno a +0,22 SD en Italia (Carlana y La Ferrara, 2025) y +0,26 SD en España (Gortazar et al., 2024)— confirman la fuerza de la tutoría bien diseñada. El arco efecto-escala se sostiene en diferentes contextos.

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating Math Recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. American Educational Research Journal, 50(2), 397–428.

Whitehurst, G. J. (2009). Don't forget curriculum. Brookings Institution.

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. Carles Aulés y Marc Colomer forman parte del Innovamat. 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 Ángela Pérez la revisión de este capítulo y su contribución a mejorar su calidad. También queremos agradecer el trabajo de maquetación de Marta Soler.*