Innovamat

Quin impacte tenen els programes de matemàtiques en l'aprenentatge?

Quanta evidència hi ha, quina millora és realista esperar i quines condicions la multipliquen.

Carles Aulés
Especialista de Recerca a l'Innovamat Research Lab

Marc Colomer
Director de Recerca de l'Innovamat Research Lab

Innovamat Research Series

Resum

Què sabem sobre l'impacte dels materials i programes de matemàtiques en l'aprenentatge? La recerca acumulada des de començaments d'aquest segle ofereix una conclusió general: **els materials curriculars poden millorar els resultats, però els seus efectes solen ser moderats i depenen en gran mesura de com s'implementen**. A més, encara sabem molt menys del que seria desitjable sobre quins programes funcionen, per a qui i en quines condicions. La revisió de l'evidència disponible permet extreure les conclusions següents:

- **L'evidència rigorosa continua sent escassa.** La major part prové dels Estats Units i, fins i tot allà, molts programes àmpliament utilitzats mai no han estat avaluats amb dissenys capaços d'estimar de forma convincent el seu impacte.

- **Obtenir evidència causal és costós i difícil.** Una avaluació rigorosa pot requerir diversos anys, milers d'estudiants, nombroses escoles, finançament considerable i centres disposats a acceptar que l'assignació dels programes es determini de forma aleatòria.

- **Quan es detecten millores, solen ser moderades, però poden ser rellevants a gran escala.** Els metaanàlisis situen l'avantatge mitjà de substituir uns materials curriculars per uns altres al voltant de +0,05 a +0,10 desviacions estàndard (SD). Entre els estudis revisats per *What Works Clearinghouse* i *Evidence for ESSA*, la mediana és de +0,13 SD. Són efectes modestos, però la comparació no és entre rebre o no rebre ensenyament: el grup de control continua fent servir els seus materials habituals. Per tant, aquestes xifres representen la millora addicional atribuïble a substituir una alternativa per una altra. Si totes dues tenen costos similars, fins i tot una millora moderada pot resultar important quan arriba a tots els alumnes.

- **Un mateix programa pot produir resultats diferents segons on, com i contra què s'avaluï.** L'impacte observat depèn de la qualitat de l'alternativa utilitzada pel grup de comparació, de la prova emprada per mesurar l'aprenentatge, del temps transcorregut des de la intervenció i de la qualitat de la implementació. Les avaluacions molt alineades amb el contingut del programa solen trobar efectes més grans que les proves externes, i part de les millores pot disminuir amb el temps.

- **Els materials funcionen millor quan van acompanyats de formació docent vinculada a la pràctica.** Els estudis sintetitzats per Pellegrini et al. (2021) i Lynch et al. (2019) troben mesures d'efecte superiors quan la intervenció combina materials curriculars amb acompanyament i desenvolupament professional (la mesura de l'efecte creix entre +0,08 i 0,10 SD). L'evidència suggereix que lliurar bons materials no n'hi ha prou: ajudar el professorat a fer-los servir bé n'augmenta l'impacte.

- **L'equilibri entre el cost i el benefici.** Intervencions intensives com la tutoria o el treball en petits grups obtenen, de mitjana, efectes alts. Una cosa semblant passa amb el *coaching*. Aquestes intervencions requereixen, però, molts més recursos humans i són difícils d'estendre. Els materials curriculars acompanyats de formació produeixen efectes més petits per alumne, però poden arribar a pràcticament totes les aules i fer-ho amb un cost addicional relativament baix.

Quan busquem recursos, materials i programes per donar suport a l'ensenyament i l'aprenentatge de les matemàtiques, l'oferta resulta aclaparadora. Hi ha llibres de text i altres materials curriculars, aplicacions digitals complementàries, eines d'avaluació, serveis de formació docent i programes integrals que combinen materials, formació i acompanyament. Són propostes molt diferents en el seu enfocament i abast. Davant d'aquesta varietat, la pregunta central és quina diferència pot produir cada opció en l'aprenentatge de l'alumnat. Importa realment triar un recurs en lloc d'un altre? Quina millora es pot esperar d'una proposta de més qualitat?

Abans de res, val la pena precisar el vocabulari que fem servir al llarg del capítol.

- Amb **materials curriculars o programa de matemàtiques**, ens referim a un conjunt estructurat de recursos educatius dissenyats per donar suport a l'ensenyament dels continguts establerts en el currículum oficial. Poden incloure llibres de text, guies per a docents, materials manipulatius, eines digitals o d'avaluació, entre altres components. Per distingir millor entre propostes de diferent abast, al llarg del capítol parlarem generalment de **materials curriculars** quan aquests s'ofereixen sols o amb un suport a la implementació mínim, i de **programa** quan els materials s'acompanyen d'un suport més substantiu a la implementació, com ara formació específica, assessorament pedagògic i/o *coaching*.

- Amb **recursos suplementaris**, ens referim a materials estructurats que donen suport a l'ensenyament de determinats continguts o habilitats, però que no pretenen cobrir el currículum de manera integral: aplicacions digitals, quaderns de pràctica, eines d'avaluació o altres materials específics.

- Amb **formació docent**, ens referim a les activitats orientades al desenvolupament dels coneixements, les habilitats i, en general, la capacitat professional dels docents.

- **Intervenció**: en el seu sentit més ampli, ens referim a qualsevol proposta deliberada per millorar l'ensenyament o l'aprenentatge de les matemàtiques. Però quan parlem d'**intervenció focalitzada**, ens referim a una cosa més concreta: la instrucció en grups petits, complementària a la classe i normalment adreçada a alumnat amb dificultats.

En qualsevol d'aquests casos, interpretar un estudi d'impacte sobre una intervenció educativa exigeix una mirada crítica, com expliquem al capítol 1 «Com sabem si un programa de matemàtiques funciona?». Ens hem de preguntar quin tipus d'estudi es va fer, amb quina alternativa es va comparar, quins resultats es van mesurar, en quin moment i en quin context. Tanmateix, valorar el rigor d'un estudi no n'hi ha prou. També necessitem un punt de referència que ens permeti interpretar la magnitud del resultat: fins i tot quan un efecte està ben estimat, és gran o petit? En termes pràctics, **quina millora és raonable esperar en adoptar un determinat programa de matemàtiques?**

Per respondre aquestes preguntes, revisem en profunditat l'evidència disponible. Abans de presentar els resultats, és útil delimitar-ne l'abast. En la nostra revisió, els Estats Units són el context en què trobem més avaluacions de l'impacte de programes de matemàtiques. En altres països i regions existeixen nombrosos estudis sobre pràctiques d'aula, trajectòries d'aprenentatge o materials concrets, però són menys freqüents les avaluacions d'un programa de matemàtiques integral. Per això, aquest capítol es basa principalment en fonts estatunidenques. Els seus resultats poden oferir un punt de partida útil per a altres sistemes educatius, però no s'han de traslladar de manera automàtica: la seva rellevància dependrà, entre altres factors, del currículum oficial, l'organització escolar, la formació docent i les característiques de l'alumnat de cada context.

La nostra anàlisi combina tres fonts d'informació. En primer lloc, examinem els principals metaanàlisis de la literatura científica. Un **metaanàlisi** sintetitza de manera sistemàtica els resultats de diversos estudis sobre una mateixa pregunta i permet estimar tant l'efecte mitjà com la seva variació. Tanmateix, les seves conclusions són tan sòlides com els estudis que inclou. En segon lloc, codifiquem 85 estudis recollits, amb data de tancament l'inici d'agost de 2026, a *What Works Clearinghouse* i *Evidence for ESSA*, dos repositoris estatunidencs que revisen l'evidència sobre intervencions educatives mitjançant criteris explícits. En tercer lloc, recopilem 229 documents que sis grans editorials estatunidenques presentaven als seus llocs web com a evidència dels seus productes. Aquests documents no són necessàriament equivalents a estudis d'impacte independents; precisament per això, analitzem quin tipus d'evidència aportava cadascun i quina era la seva solidesa.

D'aquesta revisió es desprenen **sis conclusions**, que desenvolupem al llarg del capítol. L'evidència d'impacte rigorosa és l'excepció, fins i tot als Estats Units. Generar-la és lent, car i difícil. Quan existeix, l'efecte sol ser modest, encara que pot ser una inversió eficient. Els mateixos materials curriculars no sempre tenen el mateix impacte, perquè depenen d'amb quina prova, quan i en quines condicions s'avaluen. L'impacte no depèn només de la qualitat dels materials, sinó també del que viatja amb ells: la formació, l'acompanyament docent i les condicions que permeten portar-los adequadament a l'aula. I, comparat amb altres palanques com la tutoria o el *coaching*, els materials curriculars amb formació tenen un efecte més petit per alumne però un abast molt més gran: hi ha una tensió entre impacte i escala.

Al final d'aquest capítol

sabràs quina proporció dels programes de matemàtiques compta amb evidència d'impacte sòlida i per què és tan difícil generar-la; quina millora és raonable esperar d'un canvi de programa de matemàtiques, en general modesta, encara que potencialment rellevant; per què uns mateixos materials curriculars no sempre obtenen el mateix resultat; quines condicions poden amplificar-ne l'efecte, especialment la formació i l'acompanyament docent; i quina tensió existeix entre l'impacte i l'escala de les diferents palanques que pot activar una escola.

Mida de l'efecte. La diferència entre el resultat mitjà del grup que va rebre un programa i el del grup de comparació, expressada en desviacions estàndard —SD, per les seves sigles en anglès—. En una avaluació causal ben dissenyada, aquesta diferència permet estimar quant va canviar el resultat a causa del programa.

A l'article «Com sabem si un programa de matemàtiques funciona?» expliquem amb més detall com interpretar una mida de l'efecte a nivell qualitatiu, però aquí n'hi ha prou de recordar dues idees.

La primera és que necessitem una referència realista, però no una regla universal. En la revisió ampliada de Kraft, la mediana de 3.426 estimacions procedents de 973 assaigs aleatoritzats va ser de 0,10 SD. Com a punt de partida per a estudis causals d'intervencions educatives avaluades mitjançant proves estandarditzades de rendiment, Kraft proposa considerar petits els efectes inferiors a 0,05 SD; mitjans, els situats entre 0,05 i menys de 0,20 SD; i grans, els de 0,20 SD o més. Aquestes categories ens ajuden a orientar-nos, però no substitueixen l'anàlisi del disseny, la mesura utilitzada, la població, el cost o la possibilitat d'aplicar el programa a escala (Kraft, 2020, 2023).

La segona és que una mateixa xifra en SD no representa la mateixa proporció del progrés anual típic en tots els cursos. En les dades històriques de proves estandarditzades estatunidenques analitzades per Hill et al., els guanys anuals mesurats en SD són més grans en els primers cursos i més petits a secundària. Si, com a aproximació, dividim aquest progrés anual entre nou mesos, 0,10 SD equival a prop d'un mes del progrés observat en els primers cursos de primària i a al voltant de quatre mesos entre 8è i 10è curs. Aquesta equivalència serveix per construir intuïció, però no s'ha d'interpretar com una mesura literal de mesos d'aprenentatge: depèn del curs, de la prova i de diversos supòsits sobre com es distribueix el progrés durant l'any (Hill et al., 2008; Kraft, 2020).

1 | L'evidència d'impacte rigorosa és l'excepció, no la norma

Com de comuna és l'evidència d'impacte rigorosa?

El 2023, Bellwether va examinar l'evidència disponible sobre 59 programes de matemàtiques d'editorials estatunidenques. Abans de continuar, intenta respondre aquesta pregunta: per a quants d'aquests programes creus que van trobar almenys un estudi d'eficàcia que complís uns estàndards mínims de qualitat científica?

La resposta és 17. Per als altres 42 —aproximadament set de cada deu—, Bellwether no va localitzar cap estudi que satisfés tots els seus requisits. Entre altres criteris, l'estudi havia d'haver-se fet als Estats Units, centrar-se en alumnat de K–9, incloure almenys 250 participants i haver-se publicat en algun dels formats admesos per la revisió. Cal remarcar que aquests 59 programes no es van triar de manera aleatòria. Es tracta de 59 programes que EdReports —una institució independent i sense ànim de lucre que avalua i publica revisions per experts sobre materials instructius— va identificar com de qualitat mitjana o alta. És a dir, es tracta de programes que a nivell de contingut curricular tenen un segell de qualitat. Si ampliem la llista a altres materials i programes, probablement el percentatge de programes amb estudis d'impacte disponibles seria encara més petit.

Però la història no acaba aquí. La legislació estatunidenca ESSA distingeix quatre **nivells d'evidència**. El nivell 1, anomenat evidència forta, requereix almenys un resultat favorable i estadísticament significatiu procedent d'un estudi experimental ben dissenyat i ben executat. En aquest tipus d'estudi, anomenat **assaig controlat aleatoritzat** —RCT, per les seves sigles en anglès—, l'assignació al programa o al grup de comparació es fa a l'atzar. El nivell 2, o evidència moderada, es basa en estudis quasiexperimentals rigorosos; el nivell 3, o evidència prometedora, en estudis correlacionals que incorporen controls estadístics per reduir el biaix de selecció; i el nivell 4 exigeix, com a mínim, una justificació teòrica acompanyada d'algun esforç per avaluar-ne els efectes. La pregunta és, dels estudis disponibles, quants han fet un assaig aleatoritzat alineat amb el nivell 1 d'ESSA?

Bellwether va localitzar 101 estudis d'aquests 17 programes i es va quedar amb 61 després d'aplicar els seus criteris. D'aquests, aproximadament un de cada deu feia servir un disseny aleatoritzat. La majoria dels estudis restants feia servir **dissenys quasiexperimentals**, correlacionals o altres enfocaments. És a dir, si mirem a nivell global, el país amb més recerca sobre l'impacte de programes de matemàtiques són els Estats Units. Dins dels Estats Units, la majoria d'aquests programes no tenen cap estudi que n'avaluï l'impacte. I dels que sí que en tenen, la gran majoria fan servir dissenys que no permeten afirmar amb el mateix grau de confiança que un estudi experimental que l'impacte observat es degui a la intervenció.

Per complementar aquests resultats, vam fer l'exercici invers: en lloc de partir dels estudis inclosos en catàlegs d'evidència, vam examinar què troba una escola quan entra a les **pàgines de recerca dels mateixos proveïdors**. Amb data de tancament el juliol de 2026, vam recopilar els documents que sis proveïdors amb àmplia presència al mercat estatunidenc —HMH, Curriculum Associates, McGraw Hill, Savvas, Great Minds i Amplify— presentaven als seus portals com a recerca o evidència. En total, vam codificar 229 documents associats a 31 programes de matemàtiques (Innovamat Research Lab, 2026b).

Només el 7 % d'aquests documents corresponia a assaigs aleatoritzats. Aproximadament sis de cada deu no incloïen un grup extern de comparació: presentaven, per exemple, resultats abans i després de fer servir el programa, descripcions de la seva implementació o associacions entre el grau d'ús i els resultats. Aquestes anàlisis correlacionals proporcionen informació útil i, atès que organitzar un assaig aleatoritzat és molt complex i car, és esperable que la major part del que es publica siguin anàlisis més senzilles, que també aporten informació útil sobre com s'implementa un programa i com es relaciona el seu ús amb els resultats.

Dit això, cal tenir present què pot i què no pot respondre cada tipus d'estudi. Una **anàlisi correlacional** és més senzilla, però també menys rigorosa: no permet estimar per si sola què hauria passat amb el mateix alumnat si no hagués fet servir el programa, i sense un **contrafactual** creïble observar una millora no n'hi ha prou per atribuir-la a la intervenció. A això s'hi afegeix que, el juliol de 2026, 23 dels 31 programes no comptaven amb una valoració d'impacte a *What Works Clearinghouse* o *Evidence for ESSA*. És a dir, cap institució externa no havia revisat els seus estudis per classificar-los en un nivell d'evidència. Res d'això no és un tret particular d'aquestes sis editorials. Descriu l'estat del sector en general. Generar evidència causal sobre el propi programa és lent, car i difícil.

«Generar evidència causal sobre el propi programa és lent, car i difícil.»

De 59 programes de matemàtiques estatunidencs ben valorats per EdReports, 42 no tenien ni un sol estudi d'eficàcia publicat el 2023 (Bellwether, 2023). I als webs de les sis grans editorials, la majoria d'estudis són correlacionals. En altres paraules, són estudis que no poden detectar l'efecte causal d'un programa en els resultats acadèmics dels alumnes.

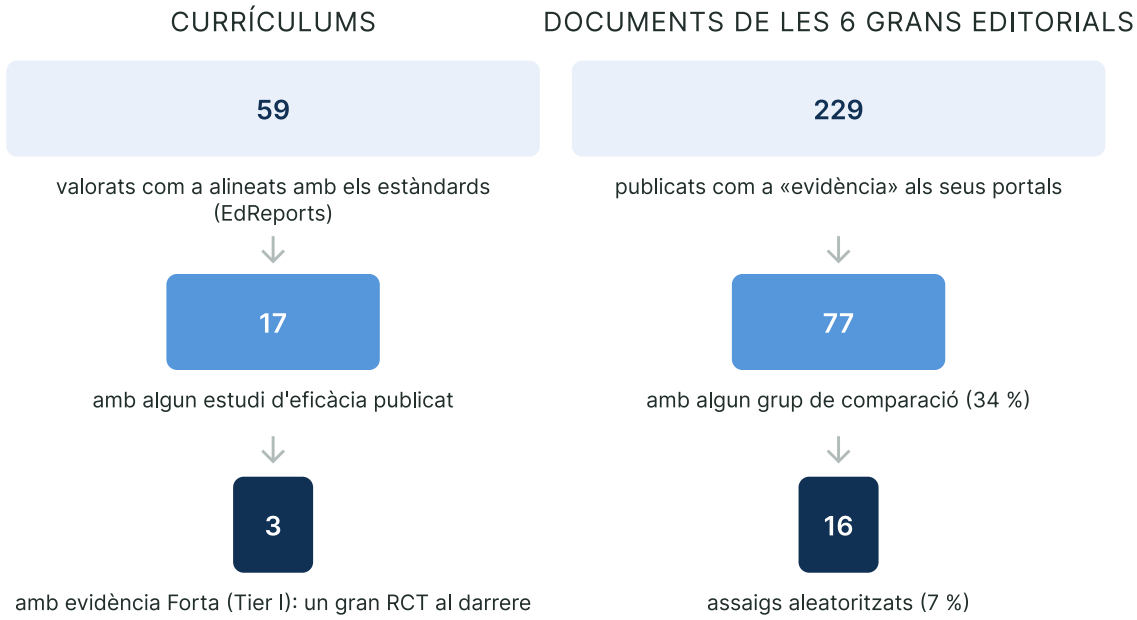


Figura 1. A l'esquerra, els programes curriculars avaluats per EdReports, segons l'estudi de Bellwether (2023). A la dreta, el que les sis grans editorials publiquen com a evidència, segons l'estudi d'Innovamat Research Lab (2026a, 2026b).

Per què hi ha tan pocs estudis amb un disseny experimental?

No hi ha una única causa. Entre les explicacions més habituals es troben les següents:

- **Cost.** Una avaluació independent a gran escala pot costar milions: la de *Cognitive Tutor Algebra I* va rebre gairebé 6 milions de dòlars de l'Institute of Education Sciences (2007). En la majoria dels casos, els estudis experimentals estan finançats per fons públics.

- **Temps.** Un assaig seriós exigeix diversos cursos entre el disseny, la implementació i l'anàlisi; el mateix estudi sobre *Cognitive Tutor Algebra I* no va trobar efectes el primer any i sí el segon (Pane et al., 2014).

- **Accés a escoles.** Cal convèncer desenes de centres i districtes que l'atzar decideixi quines escoles o aules fan servir el programa. Això implica que no tots els participants poden començar a utilitzar-lo des de l'inici, cosa que pot generar resistències i complicacions logístiques, per exemple en haver de gestionar simultàniament diferents formes de treballar. A més, poden sorgir dificultats per obtenir la col·laboració de les famílies i per evitar que els docents comparteixin materials o pràctiques entre els grups d'intervenció i control, cosa que podria reduir les diferències entre tots dos grups i dificultar l'avaluació de l'impacte real del programa.

- **Coordinació i independència.** Per augmentar la independència, convé que algú extern dissenyi, executi i publiqui l'avaluació.

- **Incentius.** Sotmetre un producte ja comercialitzat a una prova de resultat incert té pocs al·licients per a qui el ven (Bellwether, 2023).

L'evidència empírica sòlida és més l'excepció que la norma, encara que ens fixem en els Estats Units, el país amb més recerca sobre l'impacte de programes de matemàtiques. La majoria de programes de matemàtiques estatunidencs no tenen cap estudi que n'avaluï l'impacte. I dels que sí que en tenen, la gran majoria fan servir dissenys amb limitacions que no permeten concloure amb total confiança que l'impacte observat es deu al programa avaluat. Correlació no és causalitat.

2 | Quan existeix evidència, l'efecte sol ser modest

Quan un programa de matemàtiques s'avalua mitjançant un disseny capaç d'estimar-ne l'impacte, quina diferència sol produir? Quan l'avaluació es focalitza en materials curriculars concrets, en lloc d'un programa integral que inclogui també formació i acompanyament, les principals revisions apunten a una millora petita, encara que no menyspreable.

Slavin i Lake (2008) van analitzar tretze estudis sobre enfocaments curriculars —majoritàriament, llibres de text— i van estimar un efecte mitjà de +0,10 SD. Anys després, Pellegrini et al. (2021) van aplicar criteris d'inclusió més estrictes i van obtenir una estimació més petita i no significativa: +0,05 SD en trenta estudis sobre materials curriculars tradicionals i digitals que incorporaven poca formació docent. Encara que totes dues revisions difereixen en el període estudiat i en els seus criteris metodològics, ofereixen un ordre de magnitud similar. Així doncs, l'avantatge mitjà de substituir uns materials curriculars per uns altres sembla situar-se al voltant de +0,05 SD i, en algunes síntesis, prop de +0,10 SD. Aquestes xifres comparen els materials curriculars avaluats amb els que ja feia servir l'escola, que també ensenya matemàtiques, i no aïllen la qualitat dels materials: els estudis sobre adopcions reals de llibres de text troben poques diferències mitjanes entre ells encara que tinguin segells de qualitat diferents (Blazar et al., 2020).

Els estudis a gran escala sobre adopcions reals de llibres de text ofereixen una imatge igualment matisada. Blazar et al. (2020), després de combinar dades de sis estats estatunidencs després de la implantació dels estàndards Common Core, van trobar poques diferències mitjanes entre els materials analitzats. Koedel et al. (2017), en canvi, van estimar que un dels quatre llibres més utilitzats a Califòrnia superava els altres en aproximadament +0,05 a +0,08 SD a 3r de primària, amb resultats també favorables en cursos posteriors. En conjunt, aquests estudis suggereixen que les diferències entre materials curriculars poden existir, però no apareixen necessàriament en tots els contextos ni assoleixen sempre la mateixa magnitud.

L'evidència disponible no és una mostra neutral

Les xifres anteriors descriuen l'efecte mitjà que apareix en determinades síntesis de literatura científica, però l'evidència que arriba fins a nosaltres no és en absolut una mostra neutral de tots els estudis realitzats. Alguns resultats tenen més probabilitats que altres de publicar-se, incorporar-se a un catàleg d'evidència o ser destacats pels seus autors.

Aquest patró apareix amb claredat en la nostra anàlisi. Vam codificar els 85 estudis de les 52 propostes de matemàtiques del catàleg d'*Evidence for ESSA* i *What Works Clearinghouse*. Inclouen tot tipus de propostes, no només programes de matemàtiques: recursos suplementaris (aplicacions i plataformes digitals), intervencions focalitzades i tutoria, formació docent i reformes de centre; els programes de matemàtiques són 11 dels 52. Cap dels 85 estudis no presentava un efecte principal negatiu. Entre els 66 que reporten una mida de l'efecte utilitzable, la mediana és de +0,15 SD. Per tipus de programa, la mediana és de +0,13 SD tant en els 15 estudis de programes de matemàtiques com en els 27 d'aplicacions i plataformes, de +0,28 SD en les 17 intervencions focalitzades i tutories, i de +0,08 SD en els 7 de formació docent i reforma de centre (Innovamat Research Lab, 2026a). Aquestes xifres són una mica superiors a les obtingudes en diverses de les síntesis anteriors. Encara més cridaner és que no aparegui cap resultat principal negatiu.

Part d'aquesta diferència s'explica per la manera mateixa en què es construeixen els catàlegs. Els nivells 1, 2 i 3 d'ESSA exigeixen que hi hagi almenys un resultat favorable i estadísticament significatiu i que no hi hagi cap resultat negatiu rellevant. Per tant, una mostra formada a partir de programes que assoleixen aquests nivells selecciona, d'entrada, intervencions que han obtingut algun resultat positiu. No seria correcte interpretar el seu efecte mitjà com si representés el conjunt de programes de matemàtiques disponibles.

El contrast amb l'estudi de Lortie-Forgues i Inglis (2019) ajuda a apreciar aquesta selecció. Els autors van analitzar 141 grans assaigs aleatoritzats finançats amb fons públics als Estats Units i el Regne Unit,

amb més d'1,2 milions d'estudiants. Les intervencions avaluades eren molt diverses i no s'havien seleccionat per comptar prèviament amb resultats favorables. El seu efecte mitjà va ser de +0,06 SD. A més, el 40 % dels assaigs va ser no concloent: les dades no permetien inclinar clarament la balança ni cap a l'existència d'un efecte ni cap a la seva absència.

A la selecció pròpia dels catàlegs s'hi afegeix el **biaix de publicació**. Es produeix quan els estudis amb resultats positius o estadísticament significatius tenen més probabilitats de publicar-se que aquells que troben efectes nuls o negatius. En una anàlisi de 645 estudis procedents de dotze revisions de recerca educativa, Cheung i Slavin (2016) van observar que els efectes dels articles publicats eren, de mitjana, aproximadament el doble que els dels documents no publicats.

L'evidència visible pot estar condicionada, per tant, per diversos filtres relacionats, però diferents: quins estudis es publiquen, quins s'incorporen als catàlegs i quins resultats se seleccionen com a principals. Això no invalida els efectes observats, però obliga a interpretar-los amb prudència. Els +0,15 SD de la nostra base, o els +0,13 SD dels programes de matemàtiques, descriuen els estudis seleccionats en aquests catàlegs; no necessàriament l'efecte que es pot esperar de qualsevol programa de matemàtiques.

La mateixa intervenció no sempre té el mateix impacte

Com vam veure al capítol 1, la **mida de l'efecte** no és una etiqueta fixa que es pugui atribuir a un programa amb independència de les circumstàncies. A més de dependre de l'alternativa amb què es compara, varia segons què es mesura, quan es mesura i en quines condicions s'implementa. En altres paraules, no existeix un únic «efecte del programa» en abstracte, sinó efectes estimats en contextos i moments concrets.

El primer factor a tenir en compte és la **prova d'avaluació** que es fa servir. Una cosa essencial a tenir en compte quan un programa descriu la seva **teoria del canvi** és què esperem que pugui canviar o millorar, i què no. En aquest sentit, la prova utilitzada per avaluar el programa és essencial. Quan una avaluació mesura continguts, tasques o representacions molt pròxims als treballats durant la intervenció, els efectes solen ser més grans que quan s'empren proves més generals, que exigeixen aplicar el que s'ha après a situacions menys familiars (Cheung i Slavin, 2016). Fer servir proves més genèriques permet valorar millor fins a quin punt l'aprenentatge es transfereix més enllà de les activitats practicades.

L'estudi de *Fusion* n'ofereix un bon exemple. L'efecte va ser de g = 0,77 a *ProFusion*, una prova desenvolupada pels investigadors per mesurar continguts estretament relacionats amb la intervenció. En exàmens més generalistes, els efectes van ser més petits, d'entre 0,02 i 0,20 (Clarke et al., 2023). Aquesta diferència no significa que no s'aprengués res, ja que pot reflectir tant una transferència encara limitada com una prova general poc sensible al que la intervenció va treballar. Els resultats en la prova ajustada permeten saber si l'alumnat va aprendre el que la intervenció pretenia ensenyar, i les proves independents informen sobre si aquest aprenentatge es trasllada a altres tipus de tasques.

En relació amb aquest punt, en la nostra anàlisi hem observat una absència de mesures que van més enllà de l'aprenentatge (Innovamat Research Lab, 2026b), com ara les actituds cap a les matemàtiques, la identitat matemàtica o la permanència a l'escola. Els estudis publicats es focalitzen en aprenentatge mesurat a través de proves estandarditzades que, en molts casos, només capturen una part del que un programa es proposa canviar.

> «Els estudis publicats es focalitzen en aprenentatge mesurat a través de proves estandarditzades que, en molts casos, només capturen una part del que un programa es proposa canviar.»

Un segon factor a tenir en compte és el moment en què s'avaluen els resultats. Algunes intervencions produeixen un impacte positiu just en acabar que disminueix durant els mesos o anys posteriors, un fenomen conegut com a ***fade-out***. La recerca indica que aquesta pèrdua de l'impacte és freqüent en determinades intervencions adreçades a habilitats cognitives específiques, especialment quan són breus i deixen d'aplicar-se (Bailey et al., 2017, 2020). Tanmateix, no existeix una proporció universal de l'efecte que es conservi: l'evolució depèn del tipus d'intervenció, dels aprenentatges avaluats i de les experiències posteriors de l'alumnat.

El programa de tutoria individual *Math Recovery* il·lustra aquest fenomen. En acabar 1r de primària, va produir efectes d'entre +0,15 i +0,30 SD en proves administrades per avaluadors externs, i efectes encara més grans en una prova estretament vinculada al programa. Al final de 2n, en canvi, les estimacions eren petites i cap no es podia distingir estadísticament de zero (Smith et al., 2013). Això no demostra que no quedés cap aprenentatge, sinó que, un curs després, l'estudi ja no podia establir un avantatge fiable del grup que havia rebut la intervenció respecte al grup de comparació.

El tercer factor són les **condicions d'implementació**. Podem imaginar els materials curriculars com una partitura i el professorat com qui la interpreta. Sense intèrpret no hi ha música, i un mateix intèrpret no obté el mateix de qualsevol partitura. Uns materials poden facilitar tasques més riques, trajectòries d'aprenentatge més coherents, millors connexions entre idees i més oportunitats de pràctica i retroalimentació. Uns altres poden dificultar tot això. Tot i això, el resultat també depèn de com comprèn el professorat la proposta, quant de temps hi pot dedicar, com respon a les necessitats de l'alumnat i quina formació i acompanyament rep per portar-la a l'aula.

Per tant, l'efecte observat en un estudi no és únicament el resultat dels materials. És el resultat conjunt del programa, de les persones que el fan servir i de les condicions organitzatives que fan possible la seva implementació. Dues escoles poden adoptar els mateixos materials curriculars i obtenir resultats diferents si difereixen en el temps disponible, el suport al professorat, l'estabilitat dels equips o el grau en què aconsegueixen incorporar la proposta a la pràctica quotidiana. Sobre aquests aspectes en parlem més en profunditat a l'article «De què depèn aprendre matemàtiques?».

«Dues escoles poden adoptar els mateixos materials curriculars i obtenir resultats diferents si difereixen en el temps disponible, el suport al professorat, l'estabilitat dels equips o el grau en què aconsegueixen incorporar la proposta a la pràctica quotidiana.»

En resum, l'impacte d'una mateixa intervenció depèn de quatre coses: amb qui es compara (com més semblant sigui el grup de comparació i més similars siguin les activitats que fa, més difícil serà detectar un efecte positiu), amb quina prova s'avalua (una prova pròpia i alineada amb el programa donarà un efecte més gran que una d'externa i general), quan s'avalua (com més temps passa des de la intervenció, més petit sol ser l'efecte mesurat) i com s'ha implementat.

Aquesta qüestió pren una importància especial quan la intervenció s'aplica a gran escala. Un programa pot obtenir bons resultats en condicions especialment favorables —amb centres seleccionats, un seguiment estret i un suport intensiu dels seus creadors— i no reproduir-los amb la mateixa magnitud quan s'estén a contextos més diversos i funciona amb els recursos habituals del sistema. Per això, les directrius de l'Institute of Education Sciences i la National Science Foundation distingeixen entre els estudis d'**eficàcia** i els d'**efectivitat**. Els primers analitzen si una intervenció pot funcionar en condicions favorables i relativament controlades. Els segons estudien si continua produint resultats en condicions ordinàries, amb poblacions diverses i sense una participació dels creadors més gran que la que es podria esperar en una implementació habitual (IES i NSF, 2013).

Apunts finals

La idea que repeteix aquesta secció és que la mateixa intervenció no sempre té el mateix impacte: la mesura d'efecte no s'ha d'interpretar com una propietat fixa d'un programa, sinó en relació amb la pregunta que l'estudi pretén respondre: Quina diferència suposa fer servir aquest programa en lloc de l'alternativa disponible, per a un resultat concret, una determinada població i un context específic? Per això, és important recordar una cosa que comparteixen bona part dels estudis sobre materials curriculars. El grup de comparació no deixa de rebre ensenyament, sinó que continua fent servir els materials habituals. Per tant, la mida de l'efecte no representa tot el que aprèn l'alumnat amb el nou programa, sinó el seu avantatge addicional respecte a una altra proposta amb la qual també estaria aprenent matemàtiques. Aquesta comparació exigent ajuda a entendre per què les mides de l'efecte en aquests estudis se situen entre el +0,05 SD i +0,10 SD. Segons els punts de referència proposats per Kraft (2020, 2023), +0,10 SD se situa en el tram intermedi dels efectes observats en estudis causals d'intervencions educatives mesurats mitjançant proves estandarditzades.

I aquí ve la pregunta clau: és suficient aquesta millora? Val la pena la inversió? Com argumenta Whitehurst (2009) després de comparar el cost i el benefici de múltiples intervencions educatives, canviar els materials curriculars pot ser una palanca eficient, perquè els programes més eficaços no tenen necessàriament un preu d'adquisició superior al de les alternatives. Modest no vol dir irrellevant. Triar un material curricular o un altre amb prou feines canvia el cost, però pot produir una millora moderada que arriba a tot el grup.

«Modest no vol dir irrellevant. Triar un material curricular o un altre amb prou feines canvia el cost, però pot produir una millora moderada que arriba a tot el grup.»

En el millor dels casos, l'efecte típic de canviar de materials curriculars està entre +0,05 i +0,10 SD, encara que aquest efecte pot disminuir amb el temps o quan la intervenció s'implementa a gran escala. Tot i això, canviar els materials curriculars pot ser una palanca efectiva, pel fet que el cost de programes alternatius és, sovint, molt semblant.

3 | Els materials curriculars rendeixen quan no viatgen sols

Els materials curriculars rendeixen més quan arriben amb formació

Canviar els materials a l'aula de matemàtiques produeix, per si sol, un efecte petit. Com ja hem vist, el metaanàlisi de Pellegrini et al. (2021) situa l'impacte dels materials curriculars que arriben gairebé sense formació en +0,05 SD, un efecte no significatiu. Quan aquests mateixos materials van acompanyats de formació i suport a la implementació, l'efecte puja a +0,12 SD. El metaanàlisi de Lynch et al. (2019), sobre 95 estudis de matemàtiques i ciències, apunta en la mateixa direcció: els programes que incorporen només materials o només formació obtenen al voltant de +0,15 SD, mentre que els que combinen totes dues coses arriben a +0,25 SD. Els materials per si sols canvien poc l'aprenentatge. La millora apareix quan arriben acompanyats de formació específica per al professorat.

Les dues xifres no són intercanviables. La de Pellegrini (+0,12 SD) prové de set estudis de primària, només de matemàtiques, en què la formació acompanyava un llibre de text tradicional, i es compara amb els +0,04 SD d'aquests mateixos materials amb poca formació. La de Lynch (+0,25 SD) barreja matemàtiques i ciències, totes les etapes i programes que combinen formació i materials de qualsevol tipus, i inclou proves dissenyades pels mateixos investigadors, que donen efectes uns 0,27 SD més grans que les estandarditzades. El que totes dues comparteixen és la diferència: la **formació lligada als materials curriculars** afegeix entre +0,08 i +0,10 SD respecte als materials sols.

Val la pena precisar què entenen aquests metaanàlisis per formació. A Pellegrini et al. (2021), els materials curriculars «amb poca formació» són aquells en què el professorat rep menys de dos dies o 15 hores de formació, poc més que una introducció als materials; la categoria que arriba a +0,12 SD agrupa programes amb almenys dos dies o 15 hores de formació dedicada a implementar uns materials curriculars concrets a l'aula, seguida d'acompanyament (*coaching*) durant el curs. A Lynch et al. (2019), la formació inclou tallers d'estiu, sessions de seguiment durant el curs, *coaching* i treball entre docents del mateix centre, i els materials curriculars són programes que introdueixen tasques i seqüències noves amb orientacions per al docent incorporades en el mateix material. Un exemple de programa que combina totes dues coses és una unitat de fraccions acompanyada d'un taller de cinc dies a l'estiu i tretze sessions a l'escola durant el curs (Saxe et al., 2001); un altre, *SimCalc*, una unitat de proporcionalitat per a secundària amb programari i un taller de formació per al professorat, avaluada en tres assaigs a gran escala a Texas (Roschelle et al., 2010). En l'anàlisi de Lynch, la formació rendeix més quan se centra en com fer servir els materials i en com aprèn l'alumnat el contingut, quan inclou reunions de seguiment de la implementació i quan hi participen docents del mateix centre. El nombre d'hores, en canvi, no prediu l'efecte, i els components en línia s'associen a efectes més petits.

Hom podria pensar que la formació docent és el que realment crea l'impacte. Però tres grans assaigs aleatoritzats contradiuen, en part, aquesta intuïció: una formació intensiva en contingut matemàtic, però deslligada de l'aula, va aconseguir que els professors sabessin més matemàtiques sense que es veiés cap efecte positiu detectable en els seus alumnes (Garet et al., 2016); en els dos assaigs anteriors de la mateixa sèrie no va millorar ni tan sols el coneixement docent (Garet et al., 2010, 2011). L'efecte de la formació sembla arribar a l'alumnat quan canvien les pràctiques reals que es fan a l'aula de matemàtiques (Lynch et al., 2025). Aquella formació que fa pensar el docent sobre la seva pròpia aula i li dona recursos que pot portar a la pràctica té un impacte més gran que les que es limiten a prescriure conductes o afegir contingut sense exemples reals que el docent pugui adaptar a la seva aula (Kennedy, 2016).

«L'efecte de la formació sembla arribar a l'alumnat quan canvien les pràctiques reals que es fan a l'aula de matemàtiques»

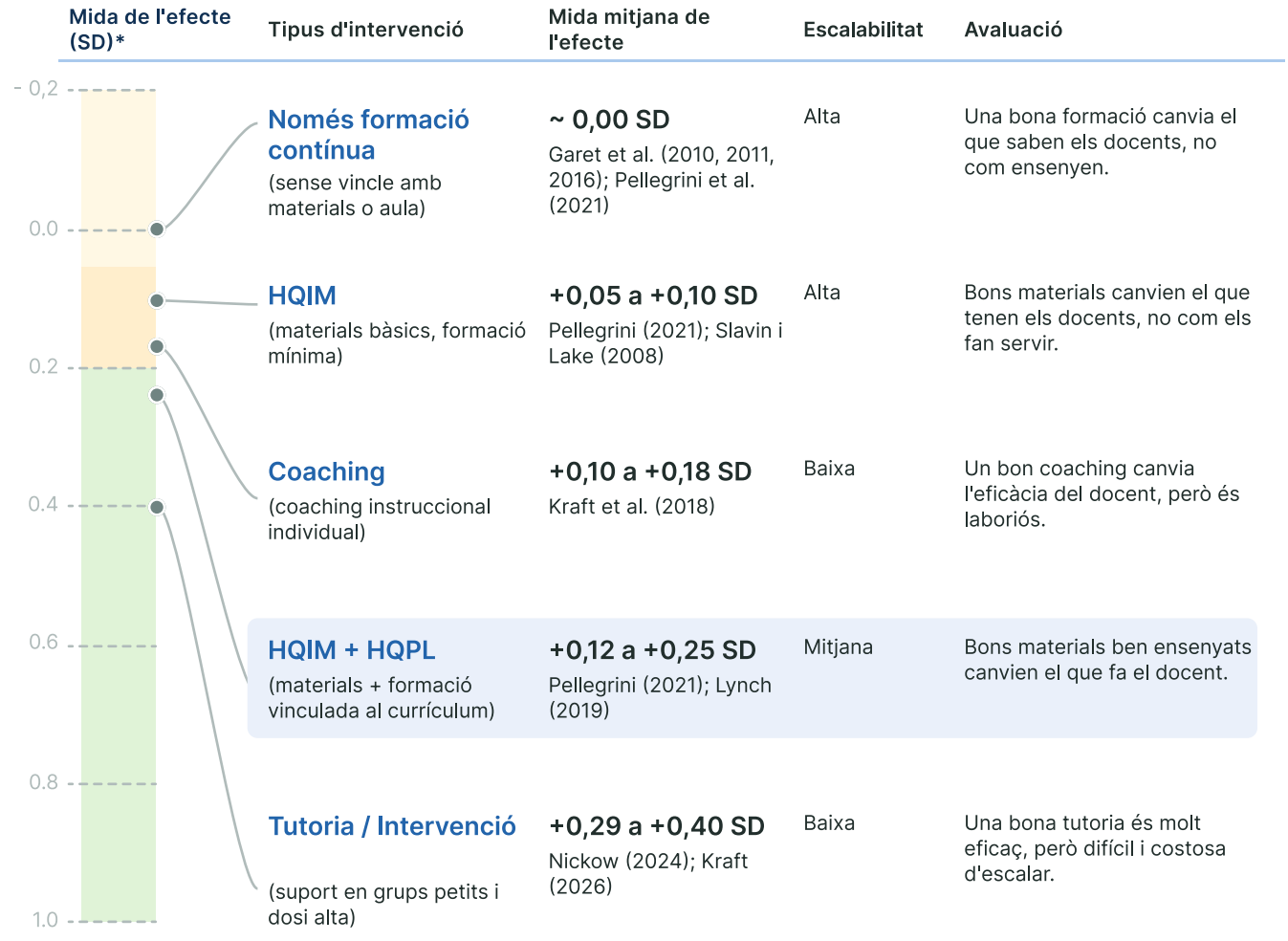
Com es compara amb altres palanques: efecte, cost i escala

Fins aquí hem comparat uns mateixos materials curriculars amb i sense formació. La segona pregunta és diferent: com es compara aquesta combinació amb altres formes de millorar l'aprenentatge que una escola pot contractar, com el *coaching* o la tutoria, en efecte, en cost i en escala?

La primera alternativa és el ***coaching***, l'acompanyament individual a l'aula d'un expert que observa i prepara classes amb el docent durant mesos (com expliquem més en detall al capítol 3). És la forma de formació docent amb més efecte conegut sobre la pràctica a l'aula: +0,49 SD (Kraft et al., 2018). El seu efecte sobre l'aprenentatge de l'alumnat és més petit, +0,18 SD, i prové gairebé del tot d'intervencions de lectura; en matemàtiques l'evidència encara és escassa i no significativa. A més, no cal veure-ho com una palanca separada dels materials: la formació lligada als materials curriculars que arriba a +0,12 SD a primària inclou aquest acompanyament com a component. Tornant a la nostra partitura: la formació general és un curset de direcció d'orquestra; el *coaching* és assajar setmana a setmana amb un mestre al costat del faristol.

Construir capacitat instruccional optimitza l'eficàcia i l'escalabilitat

Millors materials i millor suport docent es potencien mútuament. Com més es construeix el sistema per apoderar el professorat, més grans són els beneficis.



Interpretació de les mides de l'efecte (Kraft, 2020) • Petit = < 0,05 SD • Mitjà = 0,05 a < 0,20 SD • Gran = ≥ 0,20 SD

* Per claredat visual, cada intervenció se situa a la mida de l'efecte més alta observada en la literatura revisada; les estimacions més baixes dels mateixos estudis apareixen a la columna «Mida mitjana de l'efecte».

Figura 2. Comparació de la mida de l'efecte i l'escalabilitat de diferents intervencions educatives, ordenades de més petit a més gran efecte estimat. **Nota**: les plataformes digitals de pràctica queden excloses del gràfic per manca de comparabilitat: la majoria de les mides de l'efecte reportades tenen en compte els estudiants que les utilitzen amb fidelitat d'implementació, i no pas tots els estudiants de l'aula. Pel que fa al coaching, la majoria dels programes inclosos a Kraft et al. (2018) se centren en l'acompanyament docent en lectura, no en matemàtiques.

La segona alternativa, i la de més efecte directe sobre l'alumnat, és la **tutoria** en grups molt petits i dins de l'horari escolar, amb un efecte mitjà de +0,29 SD per al conjunt de programes de tutoria en un metaanàlisi de gairebé un centenar d'assaigs aleatoritzats (Nickow et al., 2024), i de +0,40 SD en la síntesi més àmplia, de 263 assaigs (Kraft et al., 2026). És la intervenció més potent que coneix la recerca educativa. Tant la tutoria com el *coaching* tenen una cosa en comú. Tenen un cost molt elevat, perquè requereixen una intervenció molt individualitzada, amb molts recursos humans i molta implicació sostinguda en el temps. Això fa que la solució sigui difícilment escalable a molts contextos. Com més s'acosta una intervenció a l'alumne concret, més costa i pitjor escala (Figura 2). La tutoria exigeix una persona per cada tres o quatre alumnes. El *coaching* exigeix un expert per cada pocs docents, i els seus efectes cauen a la meitat o més quan els programes creixen: de +0,28 SD en pilots a +0,10 a gran escala (Kraft et al., 2018). La tutoria, mesurada en implantacions grans amb tests externs, baixa a +0,16–0,22 (Kraft et al., 2026). Els materials curriculars recorren l'arc en sentit contrari: el seu efecte aïllat és el més petit, però arriba a totes les aules alhora, cada dia, a cost gairebé nul. I la formació lligada als materials, la que ensenya a fer servir els materials curriculars concrets que el docent té a l'aula, ocupa el punt d'equilibri de l'arc: augmenta l'efecte dels materials sense renunciar a la seva escala. Només la tutoria supera clarament en efecte aquesta combinació; el *coaching* la iguala, a un cost per docent molt més alt.

«Com més s'acosta una intervenció a l'alumne concret, més [recursos] costa i pitjor escala.»

El nucli pedagògic

Aturem-nos un moment a pensar què tenen en comú tots aquests resultats. La formació que funciona és la que canvia l'aula. La tutoria que funciona és la que es lliga als materials curriculars de cada dia. Els materials que funcionen són els que algú sap fer servir. Tot apunta al mateix lloc: l'aprenentatge no es produeix en el material, ni en el docent, ni en l'alumne per separat, sinó en la interacció entre tots tres. És el que la literatura defineix com a **«nucli pedagògic»** (Figura 3).

L'aprenentatge es produeix en el «nucli pedagògic»: la interacció entre docents i alumnes al voltant dels materials. Segons aquesta interacció, millorar un sol factor mentre s'ignoren els altres dos rendeix poc: l'impacte és més gran quan la intervenció o programa els té en compte tots.

Figura 3. El nucli pedagògic funciona com una interacció amb tres factors: si un factor falla, el conjunt rendeix poc. A partir de Cohen et al. (2003).

La mateixa lògica es repeteix fora dels Estats Units. El panell d'experts del Banc Mundial assenyala la **«pedagogia estructurada»** —materials de qualitat més formació i acompanyament docent, tot junt— com una de les inversions educatives més rendibles que existeixen (GEEAP, 2023). No és un programa concret: és exactament la combinació de la qual parla aquesta secció.

Fem un resum del que hem après en aquesta secció. Primer, canviar de materials curriculars pot produir efectes petits o moderats. Segon, si aquests canvis van acompanyats de formació lligada a la pràctica, les mides de l'efecte creixen. Tercer, hi ha una tensió entre **impacte i escala**: la tutoria i les intervencions en petit grup, les palanques amb més efecte, i el *coaching*, la formació més intensiva, són les que menys escalen, així que el seu paper és complementar, no substituir.

Però aquí l'analogia de l'orquestra arriba al seu límit: en una escola, el «públic» també toca. Els alumnes no reben el programa: el fan, amb el seu punt de partida, la seva família i el seu context. I al voltant de l'aula hi ha més músics dels que sembla: la direcció del centre que decideix horaris i prioritats, les polítiques educatives que fixen el currículum oficial i els recursos, els companys, les famílies. Aquest capítol ha mirat la partitura i els seus assajos; el capítol 3 obre el pla sencer: quins factors determinen l'aprenentatge de les matemàtiques, quins són modificables i quins vénen donats, i què pot fer cada actor perquè l'obra soni el millor possible (capítol 3 «De què depèn aprendre matemàtiques?»).

Els materials curriculars rendeixen entre +0,05 i +0,10 SD; amb formació docent lligada a la pràctica, entre +0,12 i +0,25 SD segons la síntesi. La tutoria i les intervencions en petit grup tenen efectes més grans, i el coaching millora sobretot la pràctica docent, però tots dos costen més i escalen pitjor: actuen com a complement, no com a solució única. L'equilibri entre efecte i escala està en els programes que combinen bons materials amb formació lligada a la pràctica.

4 | Què implica això a la pràctica

1. Ajustar les expectatives. Un efecte sostingut de 0,10–0,20 SD és del millor que ofereix aquest camp: l'alumnat arriba al mateix lloc unes setmanes o uns mesos abans. Parlem de millores rellevants però graduals, no de transformacions radicals en un sol curs. Els avenços educatius solen ser modestos, acumulatius i dependents del context, per la qual cosa les promeses de canvis espectaculars s'han d'analitzar amb cautela.

2. Demanar estudis i fer preguntes. Els estudis d'impacte i les evidències ens ajuden a prendre decisions més informades. És important demanar evidències als desenvolupadors i analitzar-les amb detall. Quina és la mida de l'efecte i com es compara amb la referència de 0,10 SD? Si els estudis no existeixen o són poc transparents amb el disseny i la mesura de l'efecte, pregunta't per què. A més, és important ser crítics quan llegim estudis, com vam veure al capítol 1: Qui signa l'estudi, el desenvolupador o un avaluador extern? Quin disseny s'ha fet servir i com s'ha estimat què hauria passat sense el programa? Amb quina prova s'ha avaluat el programa?

3. Valorar el paquet complet d'un programa, no només els materials. La formació i l'acompanyament que viatgen amb el material poden pesar més que el material mateix. En el metaanàlisi de Lynch et al. (2019), els programes que incorporen només materials o només formació obtenen al voltant de +0,15 SD, i els que combinen totes dues coses arriben a +0,25 SD. En els estudis de primària de Pellegrini et al. (2021) el contrast és equivalent: +0,05 SD gairebé sense formació i +0,12 SD amb suport a la implementació. Les dues síntesis mesuren coses diferents (primària i proves externes en un cas; matemàtiques i ciències, amb proves dels mateixos investigadors, en l'altre), però coincideixen en l'essencial: la formació lligada als materials curriculars afegeix entre +0,08 i +0,10 SD. Preguntar quin tipus de formació inclou hauria de ser tan normal com fullejar les guies per als docents o els llibres per a l'alumnat.

5 | Preguntes freqüents

«Pocs programes tenen evidència forta. Quins són els que sí?»

Tres programes de matemàtiques tenen avui la certificació més alta (evidència forta) del catàleg *Evidence for ESSA*. *Math Expressions*: +0,11 SD en el major assaig curricular públic dels EUA, més d'un centenar d'escoles assignades a l'atzar i mesurades amb un test extern (Agodini et al., 2010, 2013). *Math in Focus*: ≈+0,12 SD en un assaig aleatoritzat a Nevada, encarregat per la mateixa editorial a una consultora externa, amb efecte nul en el test estatal (Jaciw et al., 2016). I l'àlgebra de *Carnegie Learning*: efecte nul el primer any i d'aproximadament +0,20 SD el segon, en l'assaig de RAND a set estats (Pane et al., 2014).

«El programa que estem mirant no apareix a What Works Clearinghouse. Mal senyal?»

No necessàriament. Per començar, els programes no estatunidencs rarament estan llistats, simplement perquè per obtenir la certificació cal presentar estudis desenvolupats dins dels Estats Units. En aquest cas, cal mirar si tenen algun estudi publicat en una revista científica i analitzar com han definit el grup de comparació, amb quina prova es mesura l'estudi, quan i en quin context. D'altra banda, els dos grans registres estatunidencs (*What Works Clearinghouse* i *Evidence for ESSA*) se solapen molt menys del que sembla: dels 52 programes del catàleg *Evidence for ESSA*, només 3 tenen a més informe del *What Works Clearinghouse*, i a la inversa hi ha programes amb bona evidència que no són al catàleg —*Building Blocks*, a Educació Infantil, té assaigs aleatoritzats revisats per parells i no està certificat—. També hi ha absències eloqüents: el WWC va revisar *enVision*, un dels programes més venuts del país, i no va trobar ni un sol estudi que complís els seus estàndards.

«Heu analitzat les evidències de les grans editorials estatunidenques? Què hi heu trobat?»

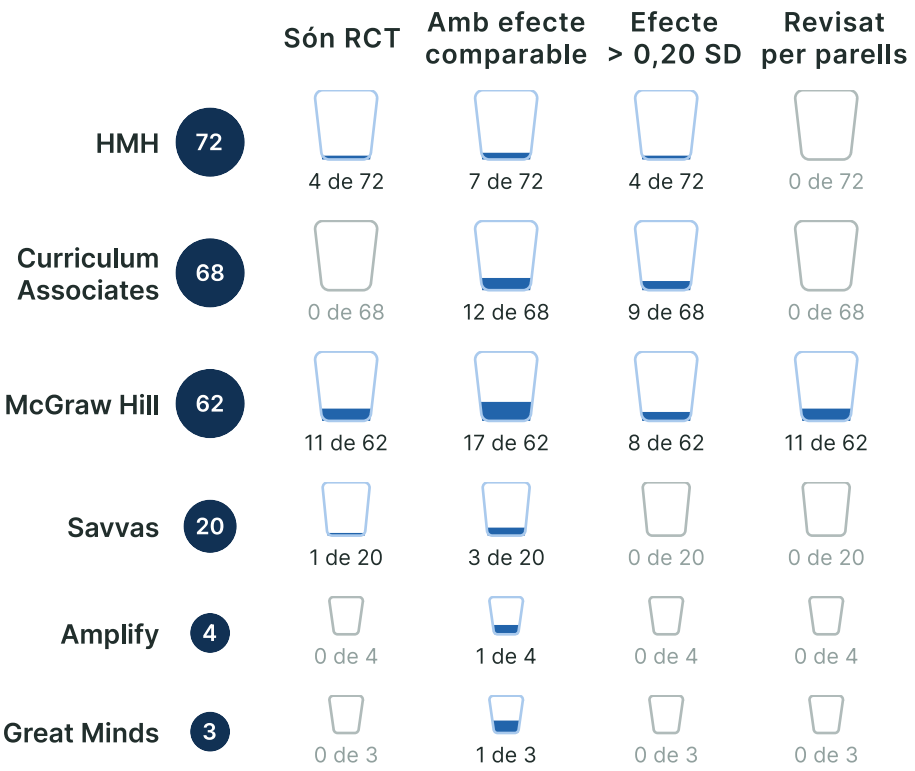
Sí, hem analitzat en profunditat les evidències públiques de sis grans editorials. Vam recopilar i codificar un a un els 229 documents que sis grans editorials presenten com a recerca o evidència en els seus propis portals (el juliol de 2026). Descriurem el tipus d'evidència que cada editorial decideix publicar al seu web, no la qualitat dels seus programes: una editorial pot tenir un bon programa i comunicar poc sobre el seu impacte, i publicar molt material no garanteix que aquest material permeti estimar un efecte.

HMH és la que més publica, amb 72 documents al seu portal de recerca, més de la meitat d'ells comparacions abans i després de fer servir el programa, sense grup de comparació. Curriculum Associates publica 68 documents, al voltant del 60 % anàlisis correlacionals, i cap no correspon a un assaig aleatoritzat. McGraw Hill en publica 62 i és l'única del grup amb articles revisats per parells, concentrats en dos programes: els assaigs aleatoritzats de *Building Blocks* van trobar efectes positius i substancials, i estan entre els estudis curriculars més sòlids publicats en revistes acadèmiques, mentre que els tres d'*ALEKS* no van trobar efectes estadísticament significatius. Savvas publica 20 documents, gairebé tots amb grup de comparació, encara que només 3 inclouen una mida de l'efecte, que és precisament el que ens permet valorar la magnitud de l'impacte i comparar-lo amb el d'altres opcions. Great Minds publica 3 documents sobre *Eureka Math* i Amplify, 4. Recordem que un volum més petit no implica per si mateix una base d'evidència més feble, ja que pot reflectir una política més selectiva sobre què es publica. En conjunt, el 7 % dels documents correspon a assaigs aleatoritzats, el 18 % reporta una mida de l'efecte estandarditzada davant d'un grup de comparació i el 5 % ha passat per revisió per parells. La base de dades completa, amb els resultats de cada estudi, està disponible de forma oberta (Innovamat Research Lab, 2026b) (Figura 4).

Val la pena posar aquestes xifres en context abans de treure'n conclusions. Un assaig aleatoritzat en condicions reals és un exercici costós i lent: exigeix convèncer desenes de centres perquè acceptin que l'atzar decideixi quin programa faran servir, sostenir l'estudi durant diversos cursos, finançar una avaluació independent i acceptar per endavant que el resultat pot no ser favorable. L'avaluació de *Cognitive Tutor Algebra I* que va fer RAND a set estats va costar prop de sis milions de dòlars i es va prolongar durant anys (Institute of Education Sciences, 2007; Pane et al., 2014). Cap editorial no pot posar en marxa un estudi així pel seu compte, i encara que volgués el cost de fer-ho és sovint astronòmic. A això s'hi afegeix que el nostre recompte descriu el que cada editorial publica al seu web en una data concreta, no tot el que existeix: hi pot haver estudis en curs, recerques no publicades o avaluacions realitzades per tercers que no apareixen en aquests portals. A més, un cop fet l'estudi, publicar-lo en revistes científiques revisades per parells encara augmenta més la lentitud i complexitat del procés i, de fet, molts estudis d'impacte vàlids mai no arriben a una revista acadèmica.

Les grans editorials: molts estudis, molt poca evidència comparable

229 documents localitzats (agost de 2026); només 41 mesuren l'impacte amb una mida d'efecte estandarditzada davant d'un grup de comparació, i només 11 han passat revisió per parells. Cada cub s'omple segons la proporció.



«Efecte comparable» = mida d'efecte estandarditzada (d o g) procedent d'un disseny amb grup de comparació; s'exclouen els pre/post sense control, els coeficients beta i els punts percentuals. A Curriculum Associates, la majoria dels efectes > 0,20 SD es mesuren amb el test de la mateixa editorial (i-Ready Diagnostic).

Font: Innovamat Research Lab (2026b); 229 documents dels portals de recerca de les sis grans editorials dels EUA, recopilats l'agost de 2026.

Figura 4. Documents sobre 31 programes de matemàtiques publicats als portals de recerca de sis grans editorials (229 documents, juliol de 2026). Les editorials publiquen molts documents i molt poca evidència comparable: només 41 dels 229 reporten una mida d'efecte estandarditzada davant d'un grup de comparació, i només 11 passen revisió per parells. A Curriculum Associates, la majoria dels efectes grans es mesuren amb el test de la mateixa editorial. Elaboració pròpia (Innovamat Research Lab, 2026b).

«No és millor un estudi de la mateixa editorial que cap estudi?»

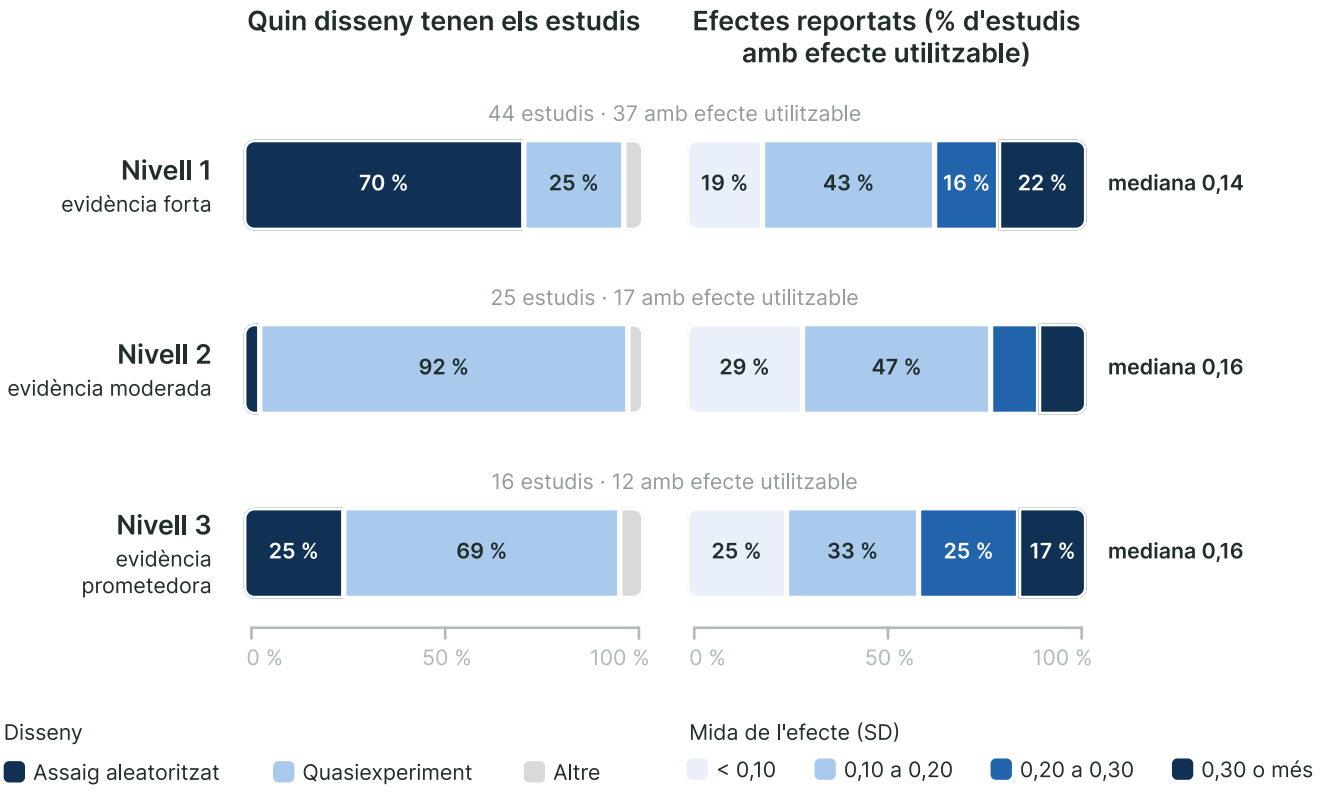
Sí, sempre que es llegeixi sabent el que és. Les dades d'algunes recerques diuen dues coses sobre els estudis interns. Primera: reporten menys i pitjor —gairebé la meitat no ofereix cap mida d'efecte, només afirmacions de millora—. Segona: tendeixen a mesurar amb una prova pròpia, ajustada al contingut treballat, i això produeix xifres sistemàticament més grans que les d'una prova externa. Els mateixos estudis ho mostren amb claredat, perquè apliquen les dues mesures al mateix alumnat i en el mateix període: *Math 180* reporta 0,51 SD en el test de la seva editorial i 0,32 en l'extern; *Math Recovery*, 1,04 en la seva mesura i 0,28 fora; *ROOTS*, 0,81 davant de 0,23; *Fusion*, 0,77 davant de 0,07, aquest últim no significatiu. Ara bé, una diferència així no demostra que no s'aprengués res: pot ser que l'aprenentatge encara no es transfereixi a tasques menys familiars, o que la prova general, molt més àmplia, no sigui prou sensible per detectar-ho. Per això cal saber sempre quina prova s'està mirant.

«Com es reparteixen els dissenys i els efectes dins del catàleg certificat?»

Hi ha una variació gran, sobretot entre nivells d'evidència. La nostra base de dades cobreix 52 programes curriculars de tota mena i 85 estudis del catàleg de matemàtiques d'*Evidence for ESSA*. El 44 % dels estudis són assaigs aleatoritzats, el 53 % quasiexperiments i un 3 % altres dissenys, i el nivell de certificació explica gairebé a la perfecció el tipus de disseny: entre els estudis de programes de nivell 1, el 70 % són aleatoritzats (no el 100 %, perquè l'etiqueta és del programa, no de l'estudi: n'hi ha prou amb un bon RCT i els altres estudis del mateix programa queden emparats sota aquesta etiqueta), mentre que en el nivell 2 el 92 % són quasiexperiments. Dels 85 estudis, 19 no reporten cap mida d'efecte utilitzable; entre els 66 restants, la mediana és +0,15 SD i els dissenys més febles reporten medianes més altes (≈0,16 en el nivell 3 davant de ≈0,14 en el nivell 1): menys rigor, xifres més grans. Les apps, el grup més nombrós del catàleg (20 de 52), s'agrupen en +0,07–0,13 SD; els efectes grans es concentren en les intervencions de petit grup i la tutoria, i un terç dels efectes de 0,25 SD o més es deu a proves alineades amb el programa (Figura 5).

Menys rigor, xifres més grans: el catàleg d'Evidence for ESSA per nivell d'evidència

85 estudis i 52 programes de matemàtiques · el tram de cada efecte es calcula sobre els estudis que reporten una mida de l'efecte utilitzable



Font: anàlisi pròpia del catàleg de matemàtiques d'Evidence for ESSA (Innovamat Research Lab, 2026a). L'etiqueta de nivell és del programa, no de l'estudi: per això hi ha quasiexperiments al nivell 1.

Figura 5. Disseny dels estudis i distribució de les mides de l'efecte del catàleg de matemàtiques d'*Evidence for ESSA*, per nivell d'evidència. Inclou tots els tipus de programa del catàleg (programes de matemàtiques, apps i plataformes, intervencions i tutoria, formació docent). Els percentatges de la dreta es calculen sobre els 66 estudis amb una mida de l'efecte utilitzable. Elaboració pròpia (Innovamat Research Lab, 2026a).

«I les apps i plataformes digitals de qualitat? Són una bona inversió?»

Les dades apunten que sí, però amb matisos. Els seus efectes (+0,07–0,13 SD) són semblants als d'un canvi de materials curriculars i el seu cost sol ser molt més baix, així que, ben triades, poden ser inversions eficients. La lletra petita: Primer, molts estudis d'aplicacions només compten l'alumnat que les va fer servir «amb fidelitat» —els usuaris assidus—, no tots els que les van rebre. Això tendeix a inflar la xifra, perquè qui fa servir molt una eina no és un alumne qualsevol, i les anàlisis

6 | Per endur-se

Tres frases resumeixen el capítol. Primer, l'evidència és l'excepció. La majoria dels programes de matemàtiques no tenen cap estudi i molt pocs tenen estudis capaços de demostrar amb confiança alta que qualsevol canvi observat es va deure a l'ús del programa. Segon, els efectes reals són modestos (al voltant d'una dècima de desviació estàndard), però sovint poden convertir-se en una bona inversió, si el cost dels productes alternatius és similar. Finalment, la palanca no són només els materials curriculars, sinó el paquet que pot oferir un bon programa: el seu efecte es més que duplica quan arriben amb formació lligada a la pràctica. El capítol 3 explica el perquè de fons: recorre els factors que determinen l'aprenentatge —de la política educativa a la família, amb el docent al centre— i assenyala quins pot moure una escola.

Analitza l'evidència d'allò que vols implementar, espera millores modestes i prioritza aquells materials curriculars amb un bon suport a la implementació.

Per aprofundir

Marcades amb ▸ les tres portes d'entrada per a qui vulgui continuar pel seu compte sense 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 compten tots els alumnes assignats troben efectes molt més petits. Segon, l'evidència més recent apunta a la part baixa del rang: el major assaig aleatoritzat recent d'una plataforma (*Zearn*, a Texas) va trobar +0,07 SD —no significatiu— en el test estatal i +0,11 en un test intermedi (Pane et al., 2025). La regla del capítol s'aplica també aquí: demanar l'efecte sobre tots els alumnes assignats, mesurat amb un test extern.

«Qui fa tots aquests estudis?»

Majoritàriament universitats i consultores de recerca. En el catàleg certificat, un 35 % els signen equips universitaris. En particular, el laboratori de Fuchs a Vanderbilt i el grup de Clarke a Oregon concentren els assaigs d'intervenció, mentre que el CRRE de Johns Hopkins treballa per encàrrec de les editorials. A part, un 40 % els signen consultores de recerca (LXD, Mathematica, SRI, AIR, RMC, Empirical Education, WestEd, RAND, HumRRO, etc.) i un 20 % els equips interns de les editorials. Només un terç ha passat revisió per parells. Als webs de les editorials la proporció s'inverteix: dos de cada tres documents els signa la mateixa casa. I una dada estructural a destacar: gairebé tots els assaigs aleatoritzats independents de programes de matemàtiques existeixen gràcies a finançament públic (sobretot, amb finançament de l'Institute of Education Sciences).

«Això passa igual fora dels Estats Units?»

L'escassetat és encara més gran. Fora dels EUA gairebé no hi ha evidència sobre programes concrets, i la recerca se centra en enfocaments pedagògics. Però les conclusions no canvien gaire. Al Regne Unit, el programa *Mathematics Mastery* va produir +0,07 SD en combinar els dos assaigs de la fundació EEF (Jerrim i Vignoles, 2016) —la xifra típica d'un canvi de materials curriculars—. En països de renda baixa i mitjana, la «pedagogia estructurada» (materials més formació i acompanyament) és una de les apostes més rendibles del panell del Banc Mundial (GEEAP, 2023). I dos assaigs europeus recents de tutoria en línia —al voltant de +0,22 SD a Itàlia (Carlana i La Ferrara, 2025) i +0,26 SD a Espanya (Gortazar et al., 2024)— confirmen la força de la tutoria ben dissenyada. L'arc efecte-escala es manté en diferents 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.

Finançament i declaració d'interessos. *Aquesta publicació ha estat finançada per Innovamat Education i forma part d'una sèrie de divulgació científica sobre educació promoguda per l'entitat. Carles Aulés i Marc Colomer formen part d'Innovamat. Les afiliacions i el finançament es fan explícits com a part del compromís de transparència de la publicació. El contingut i les conclusions són responsabilitat dels autors.*

Agraïments. *Volem agrair a l'Àngela Pérez la revisió d'aquest capítol i la seva contribució a millorar-ne la qualitat. També volem agrair la feina de maquetació de la Marta Soler.*