Innovamat

Funciona o no funciona? Com entendre l'evidència d'un programa de matemàtiques

Una guia per llegir l'evidència sobre l'eficàcia d'un programa de matemàtiques

Marc Colomer
Director de recerca de l'Innovamat Research Lab

Carles Aulés
Especialista en recerca de l'Innovamat Research Lab

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

Innovamat Research Series

Resum

Quan una escola incorpora un nou programa de matemàtiques, és habitual preguntar-se: «Funciona?». Aquest informe presenta els elements necessaris per avaluar un programa educatiu i interpretar un estudi d'impacte. S'estructura al voltant de tres idees principals:

- No és possible determinar, en abstracte, si un programa educatiu funciona. La pregunta que sí que podem fer és: «Quina diferència hi ha entre utilitzar aquest programa i utilitzar l'alternativa disponible, per a un resultat concret, una determinada població i un context específic?»

- No totes les evidències tenen el mateix pes ni aporten la mateixa informació. El rigor d'un estudi d'impacte depèn de la seva capacitat d'estimar què els hauria passat a aquests mateixos alumnes sense el programa, i això exigeix construir una comparació justa amb alumnes equivalents que no el van rebre.

- Un bon estudi d'avaluació d'impacte també requereix una bona teoria del canvi: què hauria de canviar, per mitjà de quin mecanisme i en quines condicions? Cal definir què s'espera que passi quan es fa servir el programa i considerar-ne la implementació una part fonamental del procés.

L'informe es tanca amb vuit preguntes per llegir qualsevol estudi i amb les respostes als dubtes més freqüents de qui rep un informe d'eficàcia.

Quan una escola incorpora un nou programa de matemàtiques1, apareix una pregunta recurrent entre docents i famílies: «D'acord, però... funciona?».

La pregunta és del tot legítima i cada vegada sembla més necessari avaluar si allò que fem en educació té un impacte positiu o negatiu. Així com en medicina ningú no es planteja aplicar un tractament sense que abans se n'hagin investigat els efectes, en educació aquesta idea encara no ha arrelat del tot. Afortunadament, cada vegada és més habitual avaluar l'eficàcia de les pràctiques i els recursos que entren a l'escola, i prendre decisions a partir de l'evidència disponible.

Això ens porta a un problema important. En alguns casos, la paraula «evidència» s'ha convertit en un argument de venda. Els materials arriben als centres acompanyats d'estudis, percentatges i segells que prometen millores demostrades en l'aprenentatge. És una bona notícia que cada vegada es demani més evidència empírica. Convé, això sí, saber quina pregunta respon aquesta evidència i fins a quin punt la resposta ens serveix a nosaltres.

Per començar, no podem determinar, en abstracte, si un programa *funciona*. El que sí que podem respondre és una pregunta més concreta: **quina diferència hi ha entre utilitzar aquest programa i utilitzar l'alternativa disponible, per a un resultat concret, una determinada població i un context específic?**

La manera de plantejar la pregunta és important. L'efecte d'un programa depèn d'allò amb què el comparem, de quin resultat mesurem, de qui participa en l'estudi i de les condicions en què s'aplica. Un resultat positiu ens diu què va passar en aquell context. Per saber si podem esperar un resultat semblant en el nostre, cal més informació.

Al llarg d'aquest informe veurem com es construeix una comparació que permeti estimar l'efecte d'un programa, quin paper tenen l'aleatorització i altres dissenys d'avaluació, què vol dir mesurar una millora en matemàtiques, com interpretar la mida i la precisió d'un efecte i quines condicions expliquen que un programa funcioni en uns llocs i no en uns altres. L'objectiu és aprendre a llegir un resultat d'impacte preguntant-nos **quina diferència estem estimant, com l'hem estimada i fins a quin punt podem esperar un resultat semblant en el nostre context.**

Després de llegir aquest informe

quan algú et digui que un programa de matemàtiques «funciona», sabràs quines preguntes has de fer: amb què es va comparar, què van aprendre realment els alumnes, quant van millorar i en quines condicions es va aconseguir aquest resultat. I, sobretot, podràs valorar què et diu aquesta evidència sobre el que podria passar a la teva aula.

[1] Entenem per programa de matemàtiques un conjunt de recursos educatius que combina materials concrets (llibres de text, guies per a docents, materials manipulatius, etc.) amb serveis de suport a la implementació (formació professional, assessorament, etc.).

1 | Avaluar és construir una comparació justa

Imaginem que una escola posa en marxa un nou programa de matemàtiques. Al cap d'un any, els resultats dels alumnes han millorat molt. Podem dir que el programa ha millorat el seu rendiment?

No. Els resultats poden haver canviat per moltes raons. Potser ha arribat un docent nou, el grup d'alumnes és diferent, s'han dedicat més hores a matemàtiques o l'equip docent ha rebut formació addicional. Observar que els resultats han millorat ens diu que alguna cosa ha canviat; esbrinar què ho ha causat és una altra cosa.

Aquest problema no és exclusiu de l'educació. Imaginem que una empresa farmacèutica anuncia un xarop nou i afirma que el 95 % de les persones que l'han pres s'han curat del refredat al cap d'una setmana. N'hi ha prou per concloure que el xarop funciona? Moltes persones es recuperen d'un refredat en una setmana sense prendre res. Per saber si el xarop ha tingut algun efecte necessitem saber què hauria passat a aquestes mateixes persones si no l'haguessin pres.

Posem un altre exemple. Imaginem que un estudi conclou que qui fa servir més crema solar és també qui pateix més cremades solars. És la crema la que provoca les cremades? El més probable és que aquestes persones passin moltes més hores al sol, cosa que explicaria aquesta associació.

El problema és que observar que *X* i *Y* es produeixen alhora no ens indica necessàriament que *X* hagi causat *Y*. En recerca, aquesta diferència entre associació i causalitat és fonamental, i es resumeix en una frase molt repetida: **la correlació no implica causalitat**.

> Un bon estudi d'eficàcia intenta esbrinar si fer *X* —per exemple, fer servir un programa de matemàtiques— ha causat un canvi en *Y* —per exemple, l'aprenentatge de l'alumnat—. No n'hi ha prou d'observar que les dues coses passen alhora.

Tornem a l'escola. Suposem que el nou programa s'introdueix al mateix temps que arriba un professor de matemàtiques amb molta experiència i especialment motivat. En acabar el curs, els resultats milloren. Ha funcionat el programa? Ha estat el professor? Hi han contribuït tots dos?

Amb les dades d'aquest únic curs és impossible saber-ho. Entre l'«abans» i el «després» han canviat massa coses. La comparació que realment ens interessa és una altra: com els hauria anat a aquests mateixos alumnes si no haguessin rebut el nou programa?

Aquest escenari alternatiu és el que anomenem el **contrafactual** (Gertler et al., 2016), i té una particularitat que ens planteja un problema: no el podem observar mai directament, perquè un mateix alumne no pot rebre el programa i, al mateix temps, no rebre'l.

> **Contrafactual.** Què els hauria passat a aquests mateixos alumnes si no haguessin rebut el programa o la intervenció. No es pot observar mai, i tota la credibilitat d'un estudi depèn de la qualitat de l'estimació d'aquest escenari.

Com podem estimar què hauria passat en un escenari que no s'ha produït?

Amb una comparació. Busquem un grup d'alumnes tan semblant com sigui possible al grup que va rebre el programa, però que no l'hagi rebut. El que els passi ens serveix com a referència per estimar què hauria passat al primer grup en absència del programa. Aquest és el **grup de comparació o de control**[2].

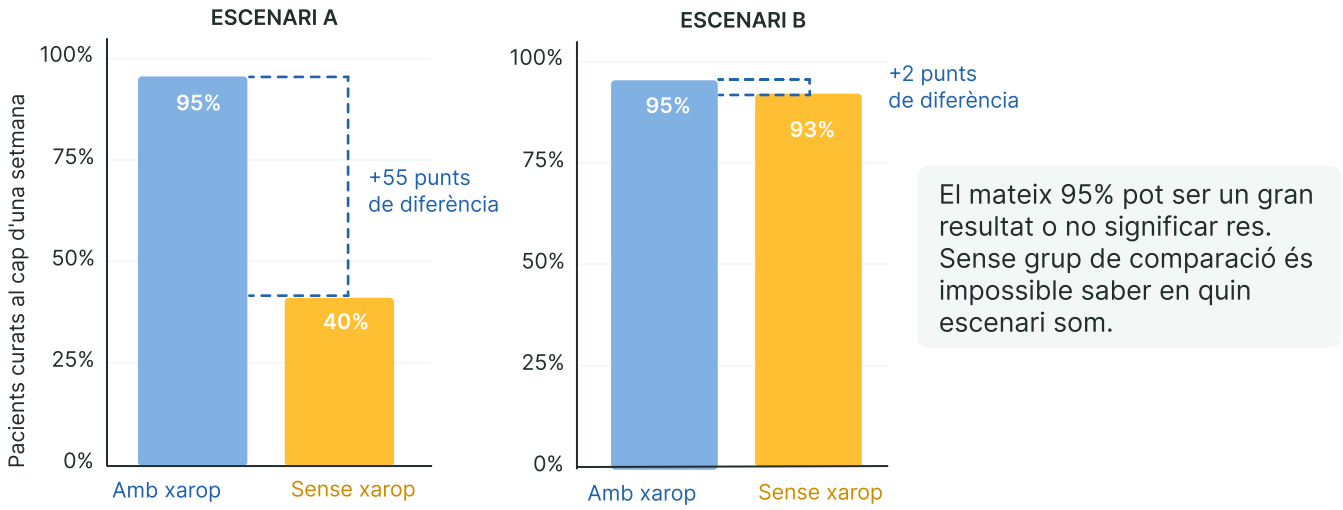
[2] Utilitzarem el terme *control* quan parlem d'estudis experimentals, mentre que *comparació* és més adequat per als estudis quasiexperimentals o correlacionals.

Tornem a l'exemple del xarop. Si el 95 % de les persones que prenen el medicament es recupera en una setmana, la xifra per si sola ens diu molt poc. Si entre persones comparables que no el prenen se'n recupera el 93 %, probablement el xarop té un efecte petit. Si només se'n recupera el 40 %, tindríem un senyal molt més clar que el medicament té un efecte. La xifra del 95 % comença a ser informativa quan tenim amb què comparar-la.

La qüestió clau és, doncs, com construïm una comparació prou justa. Quan podem decidir per sorteig qui rep el programa, aconseguim que les característiques dels dos grups siguin comparables de mitjana, fins i tot en aspectes que no sabem mesurar. Quan el sorteig no és possible, necessitem altres dissenys que construeixin una comparació igualment creïble. Aquesta és la lògica que hi ha darrere dels experiments aleatoritzats i dels dissenys quasiexperimentals.

«La qüestió clau és, doncs, com construïm una comparació prou justa.»

La comparació ens dona respostes



El mateix 95% pot ser un gran resultat o no significar res. Sense grup de comparació és impossible saber en quin escenari som.

Figura 1. Simulació de l'estudi del xarop amb grup de comparació i dos escenaris diferents. En l'escenari A, és poc habitual que els pacients es curin sense prendre el xarop. En l'escenari B, en canvi, gairebé tots els pacients (un 93 %) es curen de manera natural sense prendre res. Sense una comparació justa, no podem interpretar el valor del 95 %.

Que les notes pugin després d'adoptar un nou programa de matemàtiques encara no demostra que el programa funcioni, perquè entre l'«abans» i el «després» canvien moltes més coses que poden ajudar a explicar la millora. Avaluar l'impacte causal és estimar què els hauria passat a aquests mateixos alumnes sense el programa, i això exigeix construir una comparació justa amb alumnes equivalents que no el van rebre.

2 | Com construïm una comparació justa?

L'atzar com a eina

La manera més senzilla de construir-la és fer servir l'atzar. Imaginem que tenim un grup d'escoles interessades a provar un nou programa de matemàtiques. En lloc de deixar que cada escola decideixi si l'utilitza, sortegem quines el rebran i quines continuaran amb l'alternativa habitual. Convé fixar-se en el paper de l'atzar: decideix qui rep el programa i qui serveix de comparació entre escoles que ja havien acceptat participar-hi.

Si tenim prou escoles, l'atzar fa que, de mitjana, els dos grups siguin comparables abans de començar. Algunes escoles tindran docents més experimentats i d'altres menys, algunes tindran alumnat amb millors resultats inicials i d'altres pitjors, algunes estaran més motivades i d'altres menys. Com que l'assignació s'ha fet a l'atzar, cap d'aquestes característiques no s'hauria de concentrar sistemàticament en un dels dos grups.

I això inclou característiques que ni tan sols sabem mesurar. Podem comparar els centres segons el nivell socioeconòmic dels seus alumnes, els seus resultats anteriors o la mida de l'escola, però sempre hi haurà altres diferències més difícils d'observar, com ara les expectatives de l'equip docent, la implicació de les famílies o la cultura del centre. **El gran avantatge de l'atzar és que les equilibra de mitjana sense que ens calgui conèixer-les.** Així, si després del programa els alumnes que l'han rebut obtenen millors resultats que els del grup de comparació, tenim una raó molt més sòlida per atribuir aquesta diferència al programa.

> L'aleatorització fa que les característiques inicials de les escoles siguin, de mitjana, independents de l'assignació del programa. Les escoles continuaran sent diferents entre si; tot i així, podrem interpretar la diferència posterior entre els grups com un efecte causal amb molta més confiança.

Aquest és el principi d'un **experiment aleatoritzat**, també anomenat assaig controlat aleatoritzat (*randomized controlled trial*, RCT). És la mateixa lògica que s'utilitza en un assaig clínic, en què algunes persones reben el tractament i d'altres un placebo o l'alternativa habitual, i la diferència entre els dos grups permet estimar l'efecte del tractament. Però, en educació, l'aplicació té algunes particularitats.

Els experiments educatius presenten dificultats específiques

Moltes vegades no podem assignar el programa a l'atzar a cada alumne individualment. Si el programa modifica la manera d'ensenyar, el més habitual és fer l'assignació per aules o escoles senceres, perquè altrament un mateix docent hauria d'utilitzar mètodes diferents amb alumnes de la mateixa classe i el que passa amb uns podria afectar els altres.

A més, a diferència d'un assaig mèdic, normalment tothom sap quin programa està utilitzant. El docent sap que està ensenyant amb materials nous i l'alumnat pot saber que participa en una intervenció. Part del que observem pot deure's, doncs, a la novetat, a una atenció més gran o a l'esforç addicional que acompanya qualsevol experiment.

I hi ha una última diferència important. L'efecte d'un material curricular depèn de molt més que del material mateix. El que passa a l'aula depèn de com l'utilitza el docent, de com hi respon l'alumnat i de les condicions del centre, de manera que dues escoles poden rebre exactament el mateix programa i aplicar-lo de maneres molt diferents. En llegir els resultats d'un experiment convé recordar, per tant, què s'ha avaluat exactament. L'estudi ens diu **quina diferència va produir aquell programa, aplicat d'aquella manera, en aquella població i en aquell context, en comparació amb l'alternativa** utilitzada com a comparació.

I si no podem utilitzar l'atzar?

No sempre podem assignar un programa aleatòriament. Potser totes les escoles volen utilitzar el programa i no volen esperar, potser l'administració ja l'ha implantat en determinats centres o potser hi ha raons ètiques o pràctiques que facin impossible un sorteig.

En aquests casos els investigadors construeixen la comparació d'altres maneres, amb els anomenats **dissenys quasiexperimentals**. La lògica continua sent exactament la mateixa: trobar un grup que ens permeti aproximar-nos tant com sigui possible al que hauria passat sense el programa.

Imaginem que una administració decideix oferir un programa primer a les escoles que obtenen una puntuació inferior a un llindar determinat en una prova. Els investigadors poden comparar les escoles situades just per damunt i just per sota del llindar, si tenen raons per pensar que són prou similars i que quedar a un costat o a l'altre del límit és el que determina la participació. Altres estudis aprofiten canvis de política, implantacions progressives o regles administratives que generen grups comparables.

Un bon disseny quasiexperimental pot proporcionar evidència causal molt sòlida. Quan la regla que forma els grups és convincent i els supòsits en què es basa el disseny es comproven explícitament, la confiança que mereix s'acosta molt a la d'un experiment. El que canvia és on posem l'atenció quan el llegim. Sense la protecció que ofereix el sorteig, cal examinar amb més cura per què podem considerar comparables els dos grups i en quins supòsits es basa el disseny.

El perill de comparar grups que han decidit participar-hi

Tornem a l'exemple de les escoles que s'ofereixen voluntàriament per provar un programa. Suposem que al cap d'un any els seus resultats han millorat més que els de les escoles que no hi van participar. Ha funcionat el programa?

Potser sí, però també hi pot haver una altra explicació. Les escoles que s'hi van oferir voluntàriament poden tenir docents especialment motivats, equips directius que donen molta importància a les matemàtiques o més capacitat per introduir canvis. També podria passar el contrari: que siguin centres amb resultats especialment baixos que buscaven activament una solució.

Si aquestes característiques també afecten els resultats, la diferència que observem pot reflectir qui va decidir participar-hi. Això és el que anomenem **biaix de selecció**, i és especialment difícil de detectar perquè algunes de les diferències rellevants poden no aparèixer en les dades. Podem fer que els dos grups s'assemblin molt en mida, titularitat, nivell socioeconòmic o resultats inicials i que, tot i així, difereixin en alguna cosa important que no estem mesurant.

> Comparar els qui han rebut un programa amb els qui no l'han rebut sembla una comparació natural, però pot ser molt enganyosa. Si les persones, escoles o docents que trien participar-hi són diferents dels que decideixen no fer-ho, aquestes diferències es poden confondre amb l'efecte del programa.

La qualitat de la comparació determina la força de l'evidència

Per això, quan llegim un estudi, una de les primeres preguntes hauria de ser com es va decidir qui rebia el programa i qui no. Un sorteig ens dona una base molt sòlida; qualsevol altra regla ens obliga a entendre per què permet pensar que els grups són comparables.

Això explica per què estudis que fan servir la mateixa paraula, «evidència», permeten graus de confiança molt diferents. Un estudi pot observar que els alumnes que utilitzen un programa obtenen millors resultats; si aquests alumnes van ser seleccionats d'una manera que els fa diferents de la resta, continuem sense saber quina part d'aquesta diferència es deu al programa. Un altre estudi pot utilitzar una comparació molt més rigorosa i trobar una diferència més petita. Paradoxalment, **el segon resultat pot ser molt més informatiu**. Valorar la credibilitat de la comparació és, per tant, el pas previ a examinar la mida de l'efecte.

Una manera d'ordenar l'evidència

Aquesta diferència entre dissenys és també la raó per la qual alguns sistemes classifiquen l'evidència segons el grau de confiança amb què permet fer una afirmació causal. Un dels marcs més coneguts és el de l'Every Student Succeeds Act (ESSA, 2015) dels Estats Units, que distingeix quatre nivells: evidència forta, moderada, prometedora i fonamentació. Els nivells superiors exigeixen dissenys que permetin una inferència causal més convincent. El nivell més baix recull programes que compten amb una justificació basada en la recerca i que encara no han estat prou avaluats.

Els quatre esglaons de l'escala ESSA

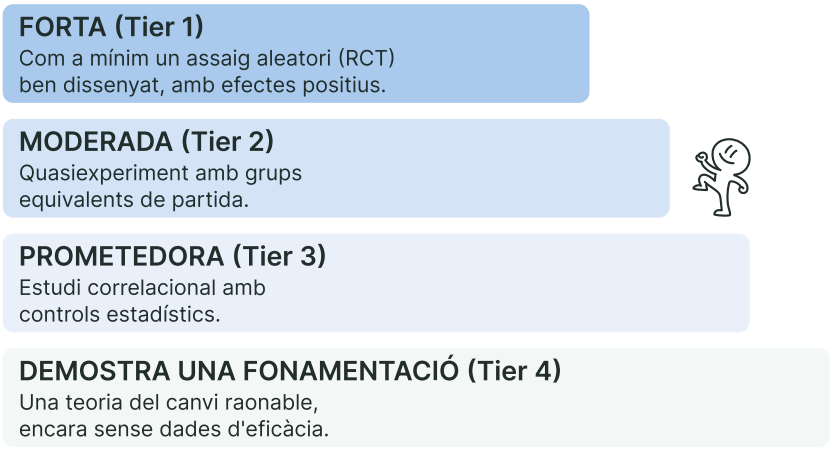


Figura 2. Els quatre nivells d'evidència de l'Every Student Succeeds Act (2015), ordenats de més a menys confiança causal.

Convé anar amb compte amb la interpretació. **Aquesta escala ordena la força de l'evidència disponible sobre els programes, i no la qualitat dels programes en si.** Un programa situat en un nivell baix pot ser prometedor i merèixer una avaluació més rigorosa. Un programa amb evidència forta compta amb estudis que permeten atribuir-li els resultats amb molta més confiança.

I ni tan sols un bon experiment no respon totes les preguntes. Ens diu si trobem una diferència entre dos grups comparables, però encara ens hem de preguntar què s'ha mesurat, quant ha millorat l'alumnat i si podem esperar un resultat semblant en el nostre context. Són les preguntes que abordarem a continuació.

La pregunta fonamental és per què podem creure que la diferència trobada es deu al programa. La resposta depèn, sobretot, de com s'hagi construït la comparació.

3 | Abans de mesurar l'impacte, cal saber què esperem que canviï

Ja sabem com construir una comparació que permeti estimar l'efecte d'un programa. Però abans cal respondre una altra pregunta: què esperem que canviï si el programa funciona?

Imaginem un nou programa de matemàtiques que proposa treballar amb més problemes oberts. Si funciona, l'alumnat tindrà més oportunitats d'explicar com ha arribat a una solució. Això podria millorar la seva capacitat de raonar i, amb el temps, aquesta millora en el raonament podria traduir-se en un millor rendiment en problemes que no ha practicat directament.

Aquesta cadena de passos és el que es coneix com a **teoria del canvi**, una explicació de com esperem que una intervenció produeixi els seus resultats. No cal que sigui complicada. En essència, ens obliga a respondre una pregunta molt senzilla: per què pensem que fer *A* hauria de produir un canvi en *B*?

Del programa a l'aprenentatge

Continuem amb el mateix exemple. En un programa que proposi als docents plantejar més problemes oberts, aquesta cadena tindria quatre etapes:

1. El programa s'implementa. Els docents reben formació i comencen a utilitzar les noves activitats.

2. Canvia la pràctica a l'aula. Els alumnes treballen realment més problemes oberts i tenen més oportunitats d'explicar les seves estratègies.

3. Canvia l'aprenentatge. Els alumnes desenvolupen millors estratègies de raonament i de resolució de problemes.

4. Canvien els resultats que ens interessen. Els alumnes obtenen millors resultats en una prova de resolució de problemes o en una mesura més general de competència matemàtica.

Cada fletxa d'aquesta cadena representa una hipòtesi que podria no complir-se. Potser el programa augmenta el nombre de problemes que es treballen però deixa intactes les estratègies que fa servir l'alumnat. Potser els alumnes desenvolupen noves estratègies però no les saben utilitzar quan s'enfronten a problemes diferents dels practicats. Potser milloren en resolució de problemes i aquest canvi és massa petit perquè es pugui detectar en una prova general de matemàtiques.

Per això, quan un estudi no detecta cap efecte, podem fer-nos una pregunta més útil que «el programa funciona o no funciona?». Podem preguntar-nos en quin punt de la cadena va deixar de produir-se el canvi esperat. Pot ser que els docents no hagin pogut aplicar el programa tal com estava previst. Pot ser que les activitats s'hagin utilitzat sense canviar realment l'experiència d'aprenentatge de l'alumnat. Pot ser que l'aprenentatge hagi canviat i que la prova utilitzada no ho detecti. O pot ser que la hipòtesi inicial fos senzillament incorrecta i que fer més problemes oberts no produeixi el tipus d'aprenentatge que esperàvem.

Mirar la cadena sencera ens ajuda a distingir entre aquestes possibilitats. I també canvia com interpretem un resultat positiu. Si trobem una millora en la prova final, podem comprovar si observem també els canvis intermedis que esperàvem. Quan apareixen, l'explicació causal resulta més coherent. Quan falten, convé ser més prudents a l'hora d'interpretar el resultat o revisar la nostra explicació de com funciona el programa.

Figura 3. Una teoria del canvi connecta la intervenció amb els resultats que esperem observar. Cada pas de la cadena planteja una hipòtesi que es pot comprovar.

Una teoria del canvi explica per què esperem que un programa produeixi determinats resultats. Ens permet passar de «aquest programa funciona» a una pregunta molt més útil: què hauria de canviar, per mitjà de quin mecanisme i en quines condicions?

Què vol dir «millorar en matemàtiques»?

La teoria del canvi ens obliga també a concretar què entenem per aprenentatge. Un programa pot estar dissenyat per millorar el càlcul, la comprensió conceptual, la resolució de problemes o la capacitat de raonar. També pot buscar canvis en la confiança de l'alumnat, la seva actitud envers les matemàtiques o la seva disposició a enfrontar-se a problemes difícils. Tots aquests aspectes poden ser rellevants, i **només sabrem si un programa ha aconseguit allò que pretenia si mesurem el resultat adequat**.

Convé, per tant, distingir entre diferents tipus de resultats.

- **Resultats propers.** Els més directament relacionats amb allò que s'ha ensenyat, com ara dominar una estratègia concreta o resoldre correctament els tipus d'exercicis que s'han practicat.

- **Transferència.** La capacitat de l'alumnat d'utilitzar allò que ha après en situacions noves, amb problemes diferents, formulacions distintes o tasques que no ha practicat directament.

- **Resultats amplis o duradors.** Canvis més generals o que es mantenen en el temps, com ara una millora global de la competència matemàtica, una confiança més gran o efectes posteriors sobre la trajectòria acadèmica.

Tots tres poden ser importants i responen a preguntes diferents. Un programa pot produir un efecte clar en un resultat proper i un efecte molt menor en una prova més general. Això pot voler dir que l'aprenentatge encara no s'ha transferit, que la intervenció actua sobre una habilitat molt concreta o simplement que estem mesurant una cosa diferent d'allò que el programa pretén canviar.

La prova forma part de la pregunta

Triar una prova és, per tant, molt més que un detall tècnic posterior al disseny de l'estudi. **La prova determina, en bona mesura, quina afirmació podem fer sobre el programa.**

Imaginem que un programa ensenya una determinada estratègia per resoldre problemes i que, al final del curs, l'alumnat obté millors resultats en exercicis pràcticament idèntics als que ha practicat. Podem concloure que ha après aquesta estratègia. Concloure que ha millorat la seva capacitat general per resoldre problemes seria molt més arriscat mentre no comprovem què passa quan les tasques són noves.

Hi ha una raó addicional per parar-hi atenció. Les proves dissenyades per qui desenvolupa un programa acostumen a estar especialment ajustades a allò que s'ensenya, i això tendeix a fer que l'efecte sembli més gran. Cheung i Slavin (2016) van revisar 645 estudis educatius i van trobar que els efectes mesurats amb proves creades pels mateixos investigadors eren aproximadament el doble de grans que els mesurats amb proves independents. Les metaanàlisis d'intervencions en ciència, tecnologia, enginyeria i matemàtiques (STEM, per les sigles en anglès) troben el mateix patró (Lynch et al., 2019).

Té sentit. Si durant mesos practiquem un determinat tipus d'exercici i després avaluem precisament aquest tipus d'exercici, estem mesurant en part fins a quin punt l'alumnat domina allò que ha practicat. Una prova externa i més àmplia planteja una exigència més gran: l'alumnat pot utilitzar allò que ha après quan canvia la tasca?

Res d'això no converteix una prova pròpia en una mala eina. Si volem saber si l'alumnat ha après exactament un contingut o una estratègia que el programa pretén ensenyar, una prova pròpia pot ser la millor opció disponible. Cada prova permet fer una afirmació diferent, i per això convé comprovar sempre qui va dissenyar la prova utilitzada i fins a quin punt s'ajustava al programa.

PREGUNTES OBERTES

- Què pretenia canviar el programa?
- Quin mecanisme hauria de produir aquest canvi?
- Quin resultat s'ha mesurat?
- La prova mesura allò que realment volem millorar?
- És una prova propera a allò que s'ha practicat o exigeix transferir allò que s'ha après?

Per interpretar un estudi d'impacte convé entendre quin canvi pretenia produir el programa, quin mecanisme havia de permetre assolir aquest resultat i quines parts d'aquesta cadena podem observar. La teoria del canvi ens diu també on mirar quan vulguem saber si el resultat es pot traslladar a una altra escola, una qüestió que reprendrem més endavant.

4 | Quant ha canviat i amb quina precisió ho sabem?

Hem vist com construir una comparació que permeti atribuir una diferència al programa i com decidir quin resultat volem mesurar. Suposem ara que fem l'estudi i trobem que els alumnes que van utilitzar el programa van obtenir millors resultats. La pregunta següent és: fins a quin punt han estat millors?

El resultat d'un estudi és sempre una estimació. Per interpretar-la, necessitem respondre dues preguntes: quina és la magnitud de la diferència i amb quina precisió l'hem estimada? La primera ens la dona la **mida de l'efecte**. La segona, l'**interval de confiança**. En els estudis també sol aparèixer el **valor p**, que ens ajuda a valorar fins a quin punt les dades són compatibles amb l'absència d'un efecte però no ens diu si la diferència trobada és gran o rellevant des del punt de vista educatiu.

Mida de l'efecte. Magnitud de la diferència entre el grup que va utilitzar el programa i el de comparació, mesurada en desviacions estàndard (SD, per les sigles en anglès). Sol expressar-se com la *d* de Cohen o la *g* de Hedges, dues mesures lleugerament diferents d'aquesta magnitud, que porten el nom dels estadístics que les van proposar.

No totes les proves utilitzen la mateixa escala. Una pot puntuar de 0 a 10, una altra de 0 a 100 i una altra pot utilitzar una escala pròpia. Per poder comparar resultats entre estudis, expressem la diferència en relació amb la variabilitat dels resultats entre alumnes, que és justament el que mesura la desviació estàndard. Un efecte de 0,20 SD significa que la diferència entre els dos grups equival a una cinquena part d'aquesta variabilitat típica.

Dit d'una altra manera, si posem els vint-i-cinc alumnes d'una classe en filera, ordenats per resultats, un efecte de 0,20 SD desplaçaria l'alumne que era al mig unes dues posicions cap amunt.

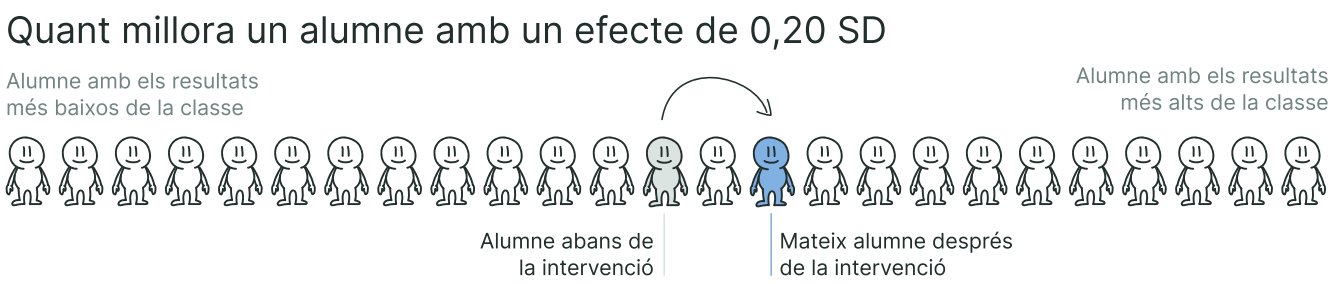


Figura 4. Els vint-i-cinc alumnes d'una classe ordenats per resultats. Un efecte de 0,20 SD desplaça l'alumne que era al mig unes dues posicions.

Trobar un efecte de 0,20 SD encara no ens diu quin és l'efecte veritable. Si repetíssim l'estudi amb altres escoles comparables, probablement obtindríem una xifra una mica diferent. L'interval de confiança expressa aquesta incertesa. Un efecte estimat de 0,20 SD amb un interval de 0,15 a 0,25 és una estimació força precisa. El mateix efecte de 0,20 SD amb un interval de −0,05 a 0,45 és molt més incert, perquè les dades resulten compatibles tant amb un efecte pràcticament nul com amb un de molt gran. El problema és més freqüent del que sembla: una revisió dels grans assajos educatius finançats als Estats Units i al Regne Unit va trobar que molts produeixen estimacions massa imprecises per aportar informació útil (Lortie-Forgues i Inglis, 2019). Per això convé llegir sempre la mida de l'efecte juntament amb el seu interval de confiança.

Una mida de l'efecte s'interpreta juntament amb el seu interval de confiança. Necessitem saber tant la magnitud estimada com la incertesa que l'envolta.

0,20 SD és molt o poc? Segons amb què ho comparem

Per interpretar una mida de l'efecte, cal una referència: hem de saber quines magnituds se solen trobar en estudis semblants.

Kraft (2020) va proposar una referència ajustada a la realitat escolar, construïda a partir de la distribució d'efectes obtinguts en assajos aleatoritzats en educació, i la va actualitzar tres anys més tard amb una mostra molt més gran. Va reunir 3.426 mides de l'efecte procedents de 973 assajos aleatoritzats i va comprovar que els efectes solen ser modestos. Més d'un terç dels efectes se situava per sota de 0,05 SD, la mediana era de 0,10 SD i només el 10 % superava 0,57 SD (Kraft, 2023). A partir d'aquesta distribució va proposar una escala específica per a intervencions educatives: menys de 0,05 SD és un efecte petit, entre 0,05 i 0,20 SD, un efecte mitjà, i a partir de 0,20 SD, un efecte gran. Una distribució molt similar va aparèixer en 138 assajos fets en països d'ingressos baixos i mitjans, també amb una mediana de 0,10 SD (Evans i Yuan, 2022).

La referència que utilitzem és molt important. En educació també s'han utilitzat altres marcs, com el de Cohen (1988), desenvolupat per a les ciències del comportament en general, o el conegut llindar de 0,40 de Hattie (2009). Tots dos han tingut molta influència i es van construir a partir de materials molt diversos, sense limitar-se a la distribució d'efectes obtinguts en experiments rigorosos fets en contextos educatius reals. Aplicar-los directament ens portaria a menystenir uns efectes que, comparats amb els que solen aconseguir les intervencions educatives, són clarament rellevants.

Per això un efecte de 0,20 SD pot semblar modest en termes absoluts i ser, alhora, un resultat important en educació. A més de preguntar-nos quina és la magnitud de l'efecte, convé plantejar-nos una altra qüestió: com és en comparació amb els efectes d'intervencions similars?

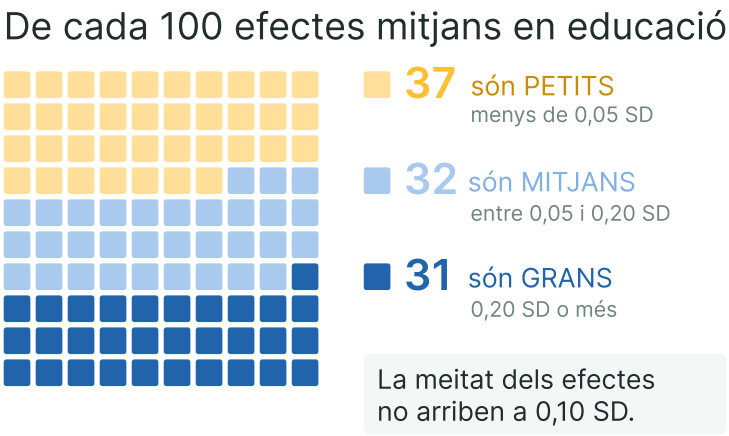


Figura 5. Distribució de 3.426 mides de l'efecte procedents de 973 assajos aleatoritzats en educació, amb la mediana i els tres trams proposats per Kraft. Elaboració pròpia a partir de Kraft (2020, 2023).

Hi ha un segon aspecte que cal tenir en compte. Una mateixa mida de l'efecte pot representar una millora més o menys rellevant segons el curs en què es mesuri.

> «Un efecte de 0,20 SD pot semblar modest en termes absoluts i ser, alhora, un resultat important en educació.»

Hill, Bloom, Black i Lipsey (2008) van estimar quant solen progressar els alumnes en proves estandarditzades al llarg d'un curs escolar i van mostrar que aquest progrés varia molt segons l'edat. En els primers cursos de primària, el creixement observat pot acostar-se a una desviació estàndard completa per any. Cap a cinquè o sisè se situa al voltant de 0,40 SD, i en els últims cursos de secundària queda per sota de 0,25 SD. En altres paraules, en els primers cursos hi ha molt marge de millora, i a mesura que els estudiants són més competents cada vegada costa més observar-hi progressos.

Així, tenim una segona referència, al costat dels efectes d'altres intervencions, per interpretar una mida de l'efecte.

Capacitat de millora d'un alumne a mesura que avança

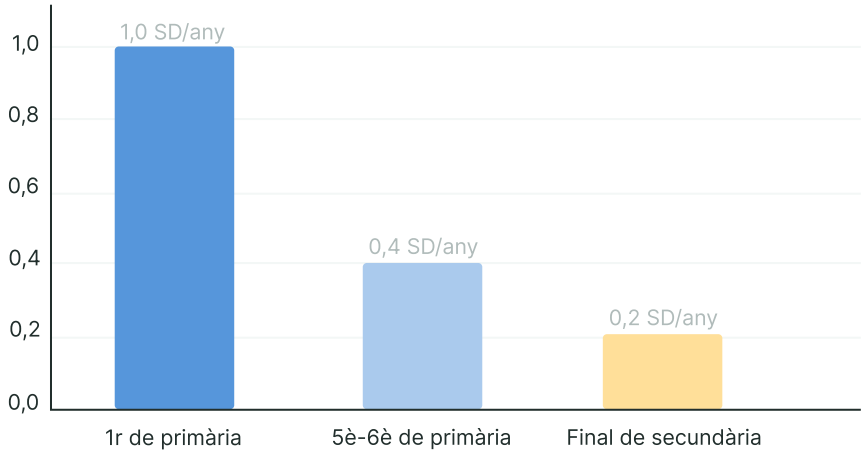


Figura 6. Creixement anual típic segons el curs, a partir de Hill et al. (2008).

Un efecte de 0,10 SD, que és l'ordre de magnitud habitual dels materials curriculars de matemàtiques (Slavin i Lake, 2008; Pellegrini et al., 2021), representa aproximadament un mes del progrés que se sol observar durant un curs en els primers anys de primària. En els últims cursos de secundària, en què el creixement anual observat és molt menor, aquests mateixos 0,10 SD representen una fracció considerablement més gran del progrés d'un curs.

De vegades aquestes comparacions s'expressen com a «mesos d'aprenentatge». Poden ser útils per fer més comprensible una mida de l'efecte però convé interpretar-les amb molta cautela. Un estudi observa una diferència en resultats i la compara amb el creixement que sol produir-se durant un curs; parlar de mesos és una traducció posterior, que depèn del curs, de la prova i de la referència utilitzada.

Mesos que triga un alumne a millorar 0,10 SD

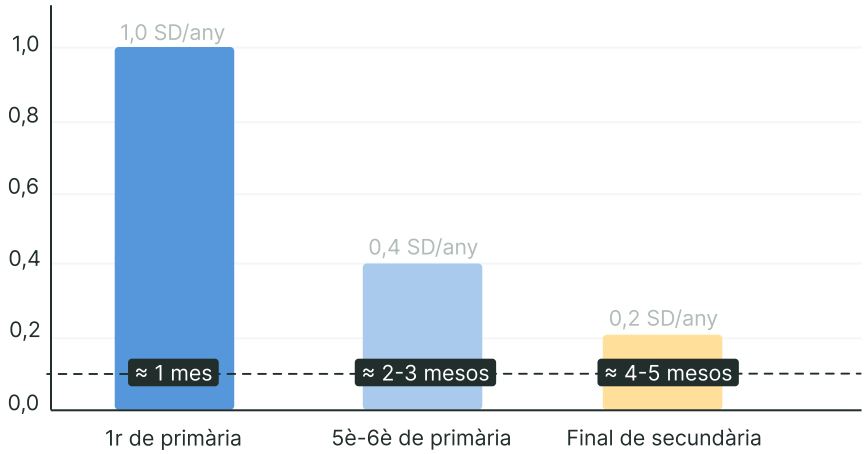


Figura 7. Estimació de quants mesos del progrés anual típic equivalen a un efecte de 0,10 SD, segons el curs, a partir de Hill et al. (2008). Aquestes equivalències són orientatives, perquè els estudis d'impacte no mesuren directament mesos d'aprenentatge.

El resultat d'un estudi és una estimació. Per interpretar-lo necessitem saber quant canvia el resultat —la mida de l'efecte— i amb quina precisió l'hem estimat —l'interval de confiança—. Després cal posar aquesta magnitud en context, comparant-la amb els efectes que solen aconseguir intervencions similars i amb el progrés habitual de l'alumnat a aquesta edat. **La mateixa xifra ens pot semblar petita o gran segons la referència que utilitzem.**

5 | Què ens diu un estudi sobre un programa?

Fins ara hem parlat d'efectes de 0,10 o 0,20 SD com si fossin una característica del programa. Convé evitar aquesta interpretació. Un estudi estima què va passar quan un programa concret es va aplicar d'una determinada manera, amb uns alumnes determinats, en comparació amb una alternativa concreta i amb una mesura específica dels resultats. Si canvien aquestes condicions, l'efecte que observem també pot canviar. De fet, una metaanàlisi que reuneix 191 experiments sobre intervencions en matemàtiques mostra que bona part de la variació entre resultats s'explica per aquestes diferències de context i de disseny (Williams et al., 2022).

Per saber què ens diu realment un estudi hem de considerar conjuntament els elements que hem anat veient, perquè la qualitat de l'evidència depèn de totes alhora. Un efecte gran resulta poc convincent si els grups no eren comparables. Un estudi molt ben dissenyat pot respondre a una pregunta poc rellevant per a nosaltres si només mesura tasques molt properes al programa. I un efecte positiu pot continuar sent incert si l'interval de confiança és molt ampli.

Fins i tot quan tot encaixa, dos estudis rigorosos sobre el mateix programa poden trobar efectes diferents sense que cap dels dos estigui equivocat. Quatre preguntes ajuden a entendre per què.

Amb què es va mesurar?

Com hem vist a la secció 3, la prova utilitzada determina en part què entenem per «funcionar». Els estudis sobre intervencions educatives solen trobar efectes més grans en proves dissenyades específicament per al programa o molt ajustades als seus continguts que en mesures més independents (Cheung i Slavin, 2016; Lynch et al., 2019). Una prova propera pot mostrar-nos si s'ha après allò que el programa ensenyava directament, i una mesura més allunyada, si aquest aprenentatge es transfereix a tasques noves. La pregunta important és quin aprenentatge representa l'efecte que estem observant.

Amb què es va comparar?

La mida de l'efecte depèn també de l'alternativa amb què es compara el programa. Comparar una intervenció amb una pràctica poc estructurada és molt diferent de comparar-la amb un altre programa de matemàtiques ben implementat. Els estudis poden estar ben dissenyats en tots dos casos i, tanmateix, respondre a preguntes diferents. Convé, per tant, saber què va rebre el grup de comparació en lloc del programa. **Com més bona sigui l'alternativa disponible, més exigent serà demostrar una millora addicional.**

Quan es va mesurar?

El moment en què es mesura també és important. Un programa pot produir una millora en acabar la intervenció que es redueixi, desaparegui o, en alguns casos, augmenti amb el temps. La revisió més completa sobre aquesta qüestió conclou que habitualment es perd una part de l'efecte, i que en moltes intervencions els estudis de seguiment detecten només una fracció de l'efecte inicial (Bailey et al., 2020). Si el que ens interessa és saber si l'aprenentatge perdura, necessitem mesures de seguiment. Un efecte observat just en acabar el programa ens diu que hi va haver una diferència en aquell moment, però no ens diu quant durarà.

En quines condicions es va aplicar?

Finalment, necessitem entendre què va caldre per obtenir el resultat. Un programa avaluat en poques escoles, amb docents especialment motivats, formació intensiva i suport continu de l'equip que el va desenvolupar pot obtenir resultats diferents quan s'estén a centenars de centres en condicions ordinàries. És un patró ben documentat. En l'acompanyament docent, per exemple, els efectes dels programes a gran escala són només una fracció dels que s'obtenen en proves pilot petites (Kraft, Blazar i Hogan, 2018).

Això no converteix un estudi petit en un mal estudi. Ens diu quines condicions van acompanyar l'efecte i ens obliga a preguntar-nos quines eren necessàries per produir-lo. Com hem vist en parlar de la teoria del canvi, la **implementació** és una de les baules que connecten el programa amb l'aprenentatge, i un bon programa pot no tenir cap efecte si no arriba a transformar de manera sostinguda el que passa a l'aula (Durlak i DuPre, 2008). Combinar materials de qualitat, formació docent i acompanyament centrat en la pràctica sol funcionar millor que aplicar aquests elements per separat (Lynch et al., 2019; Piper et al., 2018). Per això és especialment útil l'evidència que mostra que els efectes es mantenen quan un programa passa de condicions molt controlades a una implementació àmplia i semblant a la pràctica habitual.

És també la raó per la qual la comissió del National Research Council que va revisar l'avaluació de programes curriculars de matemàtiques va concloure que cap estudi aïllat no és suficient, i va proposar combinar tres perspectives complementàries: l'impacte sobre els resultats, la implementació a les aules i la qualitat matemàtica del contingut (Confrey i Stohl, 2004).

Qui ha fet l'estudi?

Convé saber també qui ha fet i finançat l'avaluació. Les revisions de la literatura educativa troben que els estudis fets o encarregats per qui desenvolupa una intervenció tendeixen a presentar efectes més grans que les avaluacions independents (Wolf et al., 2020). Els estudis petits i els que arriben a publicar-se també tendeixen a mostrar efectes més grans que els estudis amb mostres grans i que els que no es publiquen (Cheung i Slavin, 2016; Slavin i Smith, 2009; Franco, Malhotra i Simonovits, 2014).

Res d'això no invalida un estudi finançat per qui ha desenvolupat el programa, el disseny del qual s'ha d'avaluar amb els mateixos criteris que qualsevol altre. És una raó més per valorar especialment que els resultats hagin estat replicats per equips independents.

Què implica aquest resultat per a la nostra escola?

Un estudi rigorós ens diu què va passar en unes condicions determinades. Traslladar aquest resultat a una altra escola exigeix anar més enllà de buscar estudis fets en centres idèntics al nostre. La pregunta important és si en el nostre context es donen les condicions que eren necessàries perquè el programa produís el canvi.

Aquí tornem a la teoria del canvi. Si el programa va necessitar una formació docent determinada, una quantitat concreta de temps a l'aula o certes pràctiques per produir l'aprenentatge, ens hem de preguntar si aquestes condicions es poden reproduir al nostre centre. Com més depengui el resultat de condicions que nosaltres no podem reproduir, menys útil serà aquest efecte per anticipar què passaria aquí.

Per la mateixa raó, més que buscar «l'estudi» que demostra que un programa funciona, convé mirar el conjunt de l'evidència disponible. Un sol estudi pot proporcionar evidència causal molt convincent sobre el que va passar en un context concret. La nostra confiança augmenta quan trobem resultats coherents en diversos estudis rigorosos i quan aquests resultats es mantenen en canviar les poblacions, les escoles, les mesures o les condicions d'implementació. La replicació és escassa en recerca educativa, i precisament per això resulta tan valuosa quan apareix (Makel i Plucker, 2014).

Si un efecte apareix únicament amb una prova molt propera al que s'ha treballat, en un estudi petit fet amb molt de suport i just en acabar la intervenció, obtenim informació important, però limitada. Si també apareix amb mesures més allunyades, en comparació amb alternatives exigents, en implementacions a més gran escala i en contextos diferents, tenim més raons per pensar que el resultat no depèn d'una circumstància molt particular.

Un estudi ens proporciona una estimació del que va passar en unes condicions determinades. Per interpretar-la hem de considerar conjuntament la qualitat de la comparació, què es va mesurar, la magnitud i la precisió de l'efecte, amb què es va comparar el programa, quan es van mesurar els resultats i en quines condicions es va aplicar. La pregunta final és què ens permet concloure aquest estudi i quines raons tenim per esperar un resultat semblant en el nostre context. Com més coherents siguin els resultats d'estudis rigorosos amb mesures, poblacions i condicions diferents, més gran serà la nostra confiança que el resultat no depèn d'una circumstància particular.

6 | Una guia per llegir l'evidència

Al llarg d'aquest informe hem vist que decidir si un programa «funciona» requereix força més que trobar un estudi amb resultats positius. **Aquestes vuit preguntes resumeixen una manera pràctica de llegir qualsevol estudi sobre un programa educatiu.**

1. Què s'està comparant? Cal saber què reben els alumnes que utilitzen el programa i què rep el grup de comparació. L'efecte sempre és una diferència respecte d'una alternativa concreta.

2. Podem atribuir aquesta diferència al programa? Cal comprovar si hi va haver assignació aleatòria i, si no n'hi va haver, quina estratègia permet considerar que els grups eren comparables abans de començar.

3. Què pretenia canviar el programa i com? Cal conèixer-ne la teoria del canvi i identificar quins passos intermedis podem observar.

4. Què es va mesurar realment? Cal veure si la mesura capta allò que el programa ensenyava directament o si exigeix transferir l'aprenentatge a tasques noves, i si correspon al resultat que ens interessa.

5. Quant va canviar i amb quina precisió ho sabem? Cal examinar la mida de l'efecte i el seu interval de confiança, i comparar la magnitud estimada amb la d'intervencions similars i amb el progrés habitual de l'alumnat a aquesta edat.

6. En quines condicions es va obtenir el resultat? Cal saber amb quina alternativa es va comparar el programa, en quin moment es van mesurar els resultats i de quina formació, temps i suport van disposar les escoles.

7. Què sabem més enllà d'aquest estudi? Cal comprovar si hi ha altres estudis rigorosos i independents sobre el programa i si els resultats són coherents quan canvien les mesures, les poblacions i les escoles.

8. Què implica aquesta evidència per al nostre context? Cal valorar si podem reproduir les condicions que semblen necessàries perquè el programa produeixi el canvi.

La pregunta útil és què demostra un estudi, amb quin grau de confiança i fins a quin punt aquesta evidència resulta rellevant per a la decisió que volem prendre.

7 | Preguntes freqüents

«Si tots els estudis tenen limitacions, de què em serveix l'evidència?»

L'evidència no substitueix el criteri professional; ens ajuda a fonamentar-lo. En la decisió també importen les necessitats de l'alumnat, els recursos disponibles, el cost i la capacitat del centre per implementar el programa. Un efecte moderat pot ser molt valuós si el programa és assequible i fàcil d'aplicar, i molt menys interessant si exigeix una inversió desproporcionada. El que sí que pot fer l'evidència és ajudar-nos a anticipar quins resultats podem esperar raonablement, comparar alternatives, detectar promeses poc creïbles i mostrar-nos quines condicions semblen necessàries perquè un programa produeixi determinats resultats.

«Fa tres anys que treballem amb el mateix programa i els resultats han pujat clarament. Això no compta?»

Compta, i és informació valuosa sobre el que està passant a la vostra escola. La dificultat és saber quina part d'aquesta millora es deu al programa. Durant aquests tres anys poden haver canviat l'equip docent, la formació rebuda, les hores dedicades a les matemàtiques o les característiques de l'alumnat. Convé mirar també si la millora és específica del vostre centre o si acompanya una tendència general de la zona. **«Els nostres resultats han millorat des que fem servir el programa» i «el programa ha causat aquesta millora» són afirmacions diferents**, i la primera pot ser perfectament certa molt abans que tinguem prou evidència per sostenir la segona.

«El meu proveïdor m'ensenya un estudi amb un efecte de 0,45 SD. Me n'hauria d'alegrar o n'hauria de desconfiar?»

Totes dues coses, i en aquest ordre. Un efecte de 0,45 SD és més de quatre vegades superior a la mediana dels assajos educatius i supera el doble del llindar a partir del qual Kraft ja considera gran un efecte. Si l'estimació és creïble, seria un resultat excepcional. Precisament per això val la pena entendre d'on surt la xifra. Amb quina prova es va mesurar? En comparació amb quina alternativa? Quin és l'interval de confiança? Es va tornar a mesurar al cap d'un temps? Es va avaluar en poques escoles amb molt de suport o en condicions semblants a les d'una implementació habitual? Altres equips han trobat resultats semblants? **Un efecte excepcional és una bona raó per llegir l'estudi amb una atenció especial.**

«Fins a quin punt m'hauria de refiar d'un segell d'evidència?»

Un segell indica que algú ha revisat l'estudi amb criteris establerts prèviament, i això és molt útil. S'assembla al certificat de la ITV d'un cotxe, que acredita que el vehicle ha superat una revisió exigent sense garantir que el cotxe funcioni igual en qualsevol carretera. Ens informa sobre l'evidència disponible i la seva qualitat, i ens hem de continuar preguntant què es va mesurar, amb què es va comparar el programa i en quines condicions es va obtenir el resultat. També funciona a l'inrevés. **Que un programa no tingui segell pot voler dir simplement que s'utilitza en un lloc on no hi ha cap organisme que n'emeti**, o que ningú no ha presentat els seus estudis a revisió. En aquest cas haurem d'anar directament als estudis i analitzar-los a partir de les preguntes d'aquest informe.

«Qui revisa els estudis i em pot ajudar a saber si l'evidència és bona?»

Hi ha organismes especialitzats que revisen estudis educatius amb criteris comuns i en sintetitzen els resultats. Als Estats Units, el What Works Clearinghouse revisa la qualitat metodològica dels estudis (What Works Clearinghouse, 2022) i Evidence for ESSA organitza l'evidència sobre programes segons els nivells establerts per la legislació. EdReports valora la qualitat del currículum i la seva adequació als estàndards, sense entrar a avaluar l'impacte. A Europa, el referent més proper és l'Education Endowment Foundation britànica, que finança avaluacions rigoroses i sintetitza evidència educativa.

També podem recórrer a estudis publicats en revistes amb revisió d'experts, en què altres especialistes examinen els mètodes, les anàlisis i les conclusions abans de la publicació. Superar aquest procés ofereix una certa garantia que els resultats s'interpreten correctament i que les limitacions s'expliquen amb transparència. La força causal d'un estudi continua depenent, però, de com s'hagi construït la comparació i de quina pregunta permeti respondre el disseny. Els organismes de síntesi, els segells i la revisió d'experts ens ajuden a orientar-nos, però cap d'aquests recursos no substitueix les preguntes fonamentals d'aquest informe.

8 | Per acabar

La pregunta que obria l'informe, «funciona?», es pot formular d'una manera molt més útil, com ja plantejava fa vint anys la comissió del National Research Council (Confrey i Stohl, 2004): per a qui funciona, en quines condicions i per què? Un bon estudi d'impacte respon a la primera part amb una comparació creïble i una estimació honesta de la seva magnitud. La teoria del canvi ajuda a respondre les altres dues i ens permet anticipar què passarà quan el programa entri a les nostres aules.

La pregunta útil és què ha de passar perquè aquest programa funcioni aquí, amb aquests docents i aquest alumnat.

Per aprofundir-hi

El símbol ▸ assenyala dues lectures per a qui vulgui aprofundir en el tema pel seu compte sense tenir formació en recerca.

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

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

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

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

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

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

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

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

▸ Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., i Vermeersch, C. M. J. (2016). *Impact evaluation in practice* (2a ed.). Banc Interamericà de Desenvolupament i Banc Mundial.

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

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

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

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

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

Lortie-Forgues, H., i 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., i Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–303.

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

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

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

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

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

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

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

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. Marc Colomer i Carles Aulés formen part de l'equip d'Innovamat. Caterina Calsamiglia hi participa com a autora externa i hi aporta la seva experiència en avaluació d'impacte educatiu. 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 donar les gràcies a les persones que han revisat aquest informe i han contribuït a millorar-ne la qualitat: Ángela Perez, Albert Vilalta, Anna Llobet, Claudia Portas i Aida Sanuy. També volem agrair la feina de maquetació de Marta Soler i la seva contribució als gràfics.*