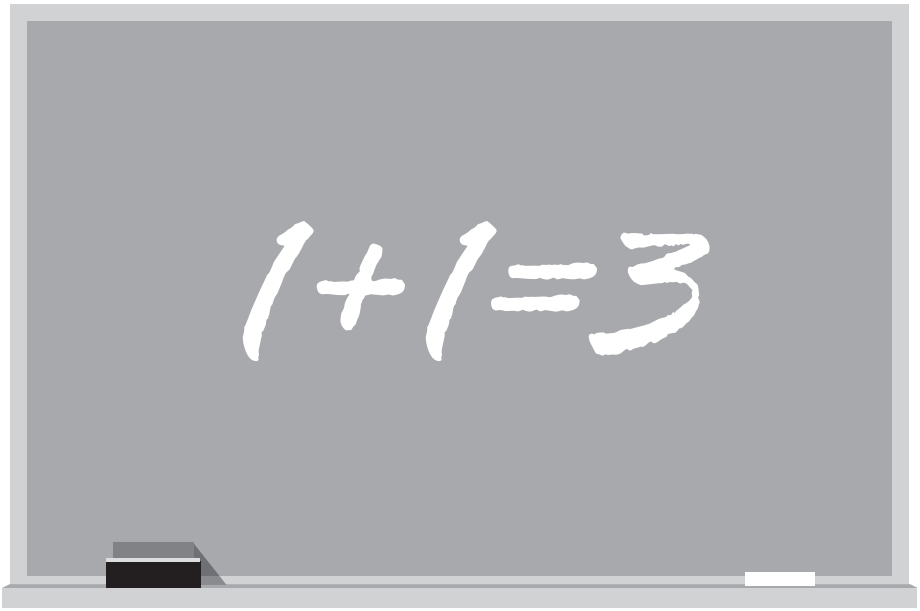
Guía práctica 11 - Cómo evaluar el impacto de las políticas educativas

Colección Ivàlua de guías prácticas sobre evaluación de políticas públicas



ivàlua | Institut Català d'Avaluació de Polítiques Públiques

Institucions membres d'Ivàlua

Generalitat de Catalunya Departament d'Economia i Coneixement
Diputació Barcelona
upf. Universitat Pompeu Fabra Barcelona
Fundació Jaume Bofill
CTESC Consell de Treball Econòmic i Social de Catalunya
Generalitat de Catalunya Consell Interuniversitari de Catalunya

Autor:
Miquel Àngel Alegre,

Maquetación y diseño portada: jaumebadosa.es

Primera edición: Abril 2015

Con la colaboración de:

Obra Social "la Caixa"

ÍNDICE

1. PRESENTACIÓN

En Cataluña ya es un lugar común la idea de que las políticas públicas deben evaluarse. El hecho de que la evaluación debe contribuir a mejorar estas políticas, sus procesos, sus resultados y, al mismo tiempo, debe servir como instrumento para rendir cuentas ante la ciudadanía. En el ámbito educativo, esta idea suele estar presente en el discurso de responsables y gestores públicos y en buena parte de la producción normativa, al tiempo que fundamenta la razón de ser de organizaciones como la Inspección Educativa o el Consejo Superior de Evaluación del Sistema Educativo.

También es cierto que en el ámbito educativo se hacen muchas evaluaciones: se examinan alumnos, se valoran méritos docentes, se diagnostican centros y se inspeccionan su gestión y sus procesos, se hace un seguimiento de indicadores educativos, se analiza la implementación y los resultados de algunos programas, etc. A todo esto lo llamamos evaluar. Es evidente, no obstante, que el tipo de conocimiento —es decir, el tipo de evidencia— que aportan las distintas prácticas evaluativas es muy diferente, como también lo es el énfasis que se ha puesto en Cataluña sobre unos tipos de evaluaciones u otros.

Actualmente disponemos de diversos instrumentos que nos permiten observar la evolución de los resultados académicos de los centros y los alumnos en distintas etapas educativas (pruebas de competencias, pruebas diagnósticas, test internacionales estandarizados, registros educativos, etc.). Por otra parte, también estamos acostumbrados a inspeccionar y documentar los procesos organizativos, la gestión académica y las actividades formativas que desarrollan los centros educativos a lo largo del curso.

En cambio, hemos dedicado pocos esfuerzos a evaluar la efectividad de los programas y las intervenciones educativas, es decir, a intentar conocer en qué medida consiguen tener un impacto sobre la realidad que pretenden modificar. En Cataluña, como en otros países de nuestro entorno, las decisiones sobre las políticas educativas —sobre su lanzamiento, su mantenimiento, su reforma, su supresión— raramente se basan en evidencias empíricas sólidas sobre su efectividad.

En cualquier escenario, más aún en el de una crisis económica y de contención del gasto, y más especialmente cuando una buena parte de nuestros indicadores educativos están todavía lejos de lo que sería deseable, es necesario conocer si las intervenciones educativas funcionan, si producen los impactos esperados. Y es necesario aprovechar este conocimiento para mejorar el diseño de las intervenciones y hacerlas más efectivas (y más coste-efectivas). Es cierto que la evaluación de impacto precisa de unas condiciones de viabilidad y de unos requisitos metodológicos no siempre fáciles de cumplir, y también es cierto que, en el proceso de formación y reforma de políticas educativas (y no solamente educativas), el recurso a la evidencia y al conocimiento deben competir, por fuerza, con otros factores (intereses o

compromisos políticos o de grupos de interés, inercias institucionales, presiones mediáticas, etc.). Pero todo ello no justifica el poco peso que tienen en Cataluña la producción y el uso de evidencias de efectividad en educación.

El propósito final de esta guía es contribuir a contrapesar esta tendència[1]. Con esta finalidad, cubre tres centros de atención.

En primer lugar (capítulo 2), discutiremos sobre el concepto en sí de evidencia en educación. Nos preguntaremos qué es evidencia sólida y creíble y qué no lo es; en qué pruebas empíricas podemos confiar y en cuáles no. Una vez hecha esta distinción, prestaremos atención brevemente a las aproximaciones y a los métodos susceptibles de generar el tipo de evidencia por el que apostaremos: el método experimental (estudios con asignación aleatoria) y determinadas técnicas cuasiexperimentales (en particular, las que permiten identificar grupos de comparación válidos).

En segundo lugar (capítulos 3 y 4), la guía ofrece ejemplos de evaluaciones realizadas y evidencias acumuladas a nivel internacional en distintos ámbitos de la política y la intervención educativa. La selección de ejemplos que aquí proponemos no tiene ningún ánimo de exhaustividad ni pretende ser representativa del conjunto de ámbitos de actuación relevantes en educación. Esta selección obedece, en cambio, a la voluntad de ofrecer una muestra lo bastante amplia y equilibrada de ejemplos de prácticas y métodos de evaluación (experimentales y cuasiexperimentales), que conecte tanto con los puntos calientes del debate público como con la agenda de prioridades de la política educativa en Cataluña.

En tercer lugar (capítulo 5), nos fijaremos en los espacios de oportunidad que podríamos encontrar en Cataluña de cara a favorecer la evaluación de impacto de las políticas y las intervenciones educativas y para promover decisiones en educación más basadas en la evidencia.

Notas:

[1] *Compartiendo este mismo propósito, debemos hacer referencia a la puesta en marcha del proyecto «Qué funciona en educación», liderado por Ivàlua y la Fundación Jaume Bofill. Este proyecto, que comenzó en el año 2015, incluye una publicación periódica de revisión de evidencias y todo un conjunto de seminarios y jornadas dirigidos a promover el aprovechamiento de este conocimiento para la mejora de la intervención y la práctica educativa. El proyecto se inspira en dos iniciativas de institucionalización de la perspectiva «qué funciona» en educación: el What Works Clearinghouse (WWC), nacido en el año 2002 de la mano del Institute of Education Sciences del gobierno federal de los Estados Unidos (http://ies.ed.gov/ncee/wwc/); y la Education Endowment Foundation, constituida el año 2011 en el Reino Unido (http://educationendowmentfoundation.org.uk/). Una parte esencial de la tarea de estas dos instituciones pasa por revisar, calificar, resumir y difundir el conocimiento existente sobre la efectividad de distintas líneas de actuación educativa, tomando como base la evidencia aportada por estudios y revisiones sistemáticas de especial relevancia.*

2. QUÉ FUNCIONA EN EDUCACIÓN: CÓMO SABERLO

Nos preguntamos, por tanto, por la efectividad de las intervenciones educativas, por sus impactos, por si funcionan o no a la hora de incidir en la realidad a la que van dirigidas. Y partimos de la base de que esta pregunta requiere respuestas basadas en evidencias empíricas sólidas, evidencias que establecen de forma robusta las relaciones de causalidad entre la política y eventuales cambios en la problemática en cuestión. Llegados a este punto, sería necesario preguntarse: ¿qué deberíamos considerar evidencia empírica creíble y sólida en el campo educativo? ¿Qué tipo de evidencia necesitamos obtener o acumular para determinar con garantías si una política o un programa educativo funciona o no funciona, si es más o menos efectivo? Preguntémonos, primero, en qué «evidencia» no deberíamos confiar.

2.1 NI INTUICIÓN, NI EVIDENCIA ANECDÓTICA

En primer lugar, es necesario abundar en el mensaje de que la intuición no es suficiente. De hecho, cuando no va acompañada de un contraste empírico fiable, la intuición puede llegar a ser un mal compañero de viaje, una mala consejera en el proceso de producción y cambio de las políticas públicas. Por ejemplo, un programa de digitalización de las aulas que proporcione a cada alumno un ordenador personal, aumentar el número de horas que pasan los alumnos diariamente en el colegio, hacer repetir curso a partir de un determinado umbral de asignaturas suspendidas, trabajar con grupos-clase reducidos de nivel homogéneo, incentivar económicamente al profesorado, permitir que las familias tengan más capacidad de elección del colegio que quieren para sus hijos o que los colegios tengan mayor libertad para definir los contenidos del currículo, etc.; todas ellas pueden parecer apuestas razonables, de las cuales podríamos esperar efectos positivos para el progreso educativo de los alumnos. Muy probablemente esto será lo que nos dicte la intuición. Pues bien, tendremos ocasión de constatar cómo la evidencia internacional no es concluyente en cuanto a la capacidad de impacto de buena parte de estas medidas, lo que equivale a decir que la evidencia empírica es, o puede a menudo ser, contraintuitiva.

En segundo lugar, deberíamos desconfiar también de lo que podríamos denominar evidencia anecdótica o circunstancial. Con este término nos referimos a aquellos datos empíricos obtenidos sin ninguna intención ni criterio de validez o representatividad, en general basados en el conocimiento personal directo (experiencia propia) o indirecto (conocimiento típico o del boca-oreja). En efecto, podemos conocer casos de alumnos repetidores que, gracias al hecho de repetir, han podido tomar conciencia de su situación académica y han conseguido reconducir su trayectoria educativa, o casos de alumnos que relajan su nivel de esfuerzo al recibir ayudas económicas a la continuidad escolar, o casos de alumnos vulnerables (académica y socialmente) que, al compartir aula con compañeros de otros niveles académicos y entornos sociales, experimentan procesos de desplazamiento y acaban abandonando los estudios, o casos de alumnos también vulnerables que rinden poco en la primaria a pesar de haber sido

escolarizados antes de los tres años de edad. Podemos haber conocido estas situaciones en primera persona o como familiares implicados, o bien haber oído hablar de ellas. Sin embargo, elevar el dato anecdótico (a menudo cuestionable por sí mismo y necesariamente tendencioso) a la categoría de evidencia es incorrecto desde cualquier punto de vista.

2.2 CORRELACIÓN NO ES CAUSALIDAD: EL EJEMPLO (FICTICIO) DEL PAE

El que las medidas de dos variables estén correlacionadas no implica que entre una y otra exista una relación de causalidad. Esta máxima, que en el plano teórico es incontestable, en la práctica no siempre resulta tan obvia. Veámoslo a través de un ejemplo.

Imaginemos que queremos conocer la efectividad de un nuevo programa de mejora de los aprendizajes instrumentales (catalán, castellano y matemáticas) dirigido al alumnado que comienza la ESO con carencias significativas en estas áreas. Lo denominaremos Programa de Aceleración Educativa (PAE). A través de un trabajo en grupos reducidos, que se prolonga durante una parte significativa del horario lectivo, el programa proporciona a los alumnos que participan en él un número de horas de clase en materias instrumentales superior al del resto de alumnos. Todo ello aunado al objetivo de reforzar y mejorar a corto y medio plazo el progreso educativo de este alumnado académicamente vulnerable. La participación en el programa, tanto por parte de los centros educativos como por parte de los alumnos, es de carácter voluntario. En el caso de los centros, la decisión de solicitar el programa debe ser aprobada por una amplia mayoría del consejo escolar. En el caso de los alumnos, la participación en el programa debe ser solicitada por la familia[2].

Por tanto, queremos saber si el PAE («tratamiento» en cuestión) puede tener impactos significativos sobre el rendimiento de los alumnos en las áreas curriculares abordadas y sobre sus niveles de abandono prematuro. Imaginémonos que, con este objetivo, comenzamos a buscar la presencia de posibles asociaciones significativas entre el hecho de participar en el programa y determinados resultados (o outcomes) como, por ejemplo, el rendimiento académico y los niveles de graduación. Y digamos que, efectivamente, identificamos la existencia de asociaciones significativas a tres niveles.

En el plano de la comparativa internacional, observamos que el alumnado académicamente más vulnerable tiende, al iniciar la secundaria, a obtener mejores resultados (en rendimiento y graduación) en aquellos sistemas educativos donde se apuesta de manera decidida por programas comparables al PAE que en otros sistemas donde este tipo de intervenciones son residuales. Veríamos, por ejemplo, que a medida que los países aumentan el porcentaje de alumnos inscritos en colegios con un PAE (o similar), aumentan también el rendimiento medio y los niveles de acreditación de su alumnado más vulnerable. Y podría ser perfectamente que esta asociación siguiera siendo significativa a igualdad de otras variables de contexto que pudiéramos considerar relevantes (PIB per cápita de los países, niveles de gasto en educación,

volumen total de alumnos con carencias formativas en primaria y secundaria, etc.).

En un segundo nivel, situados ya en Cataluña, podríamos detectar que es justamente en los centros con PAE donde el alumnado vulnerable obtiene, de media, mejores resultados. Y podría ser también que esta asociación entre PAE y el progreso académico siguiera siendo significativa a igualdad de otras variables clave de centro (porcentaje de alumnado vulnerable en el colegio, recursos materiales y humanos, ratio media de alumnos-clase en los grupos ordinarios, titularidad, etc.).

Finalmente, en un tercer nivel, podríamos observar que, dentro del mismo centro, los jóvenes vulnerables que reciben el PAE tienden a presentar unos niveles de rendimiento y graduación superiores a los jóvenes vulnerables que no solicitaron participar en él; una tendencia que podría seguir siendo significativa incluso al tener en cuenta otras variables individuales previsiblemente relacionadas con los outcomes considerados (variables como el sexo, la procedencia, el estatus socioeconómico de padres y madres, etc.).

Identificadas estas tres asociaciones, ¿podríamos entonces concluir que el PAE tiene impactos positivos sobre las oportunidades educativas del alumnado más vulnerable? Todo apuntaría hacia una respuesta afirmativa. En términos de resultados académicos, los países que apuestan por el PAE van mejor, los colegios que lo tienen van mejor, los alumnos que lo reciben van mejor. Y, en cambio, la respuesta al interrogante planteado sería: «No necesariamente». Correlación no es causalidad, y los indicios expuestos hasta aquí son eminentemente correlacionales.

En el nivel de la comparativa entre países, son muchos los factores que pueden estar mediando en la asociación entre nivel de extensión del PAE y resultados educativos, factores no siempre fáciles de controlar y que, sin embargo, pueden estar relacionados tanto con la variable de tratamiento como con los outcomes de interés; tendríamos aquí lo que se denomina un problema de endogeneidad[3]. Por ejemplo, puede ser que programas como el PAE tengan tendencia a consolidarse en sistemas educativos poco segregados, más comprensivos o más maduros en la aplicación de medidas efectivas de atención a la diversidad, o incluso en sociedades particularmente homogéneas desde un punto de vista socioeconómico y cultural. Y podría ser que fueran estas variables las verdaderas facilitadoras de la recuperación de los logros formativos de los alumnos más vulnerables.

En el nivel de la comparativa entre colegios y entre alumnos dentro de los colegios, hay que contar con los problemas de endogeneidad procedentes de posibles sesgos de selección[4]; sesgos en el acceso al programa que acaban induciendo a comparaciones entre colegios (tratados y no tratados) y entre alumnos (tratados y no tratados) muy dispares en determinadas características. Esto es particularmente evidente en el caso de programas donde la participación es voluntaria, como es el caso del PAE. En lo que respecta a los colegios,

incluso cuando comparamos centros similares en lo observable (características tangibles y mensurables), deberíamos contar con que un centro con PAE seguramente será diferente a otro sin PAE en aspectos que van más allá del hecho de disponer o no del programa en cuestión, por ejemplo, en culturas institucionales y dinámicas organizativas estrechamente relacionadas con la voluntad de acceder al programa y con la capacidad de sacarle provecho. Igualmente, las familias y alumnos que apuestan por el PAE probablemente se diferencien de otras que, pudiendo hacerlo, no lo hacen, en aspectos estratégicos y motivacionales claramente asociados con su rendimiento posterior; aspectos no observables, intangibles, que el control estadístico difícilmente puede capturar.

2.3 LOS LÍMITES DEL «ANTES-DESPUÉS»

Tampoco podemos fiarnos de los resultados de la simple comparativa entre cómo se presentaba la realidad en cuestión antes de la implementación del programa y cómo se presenta después. Por una cuestión muy clara: la dimensión sobre la cual una determinada intervención pretende generar un impacto está sometida a la influencia de otros factores externos además de la intervención objeto de evaluación. Así, los cambios que se observan en los outcomes de interés cuando ya ha terminado el programa o intervención no tienen por qué ser necesariamente efecto de su implementación. Volvamos al ejemplo anterior.

Imaginemos que identificamos un escenario preprograma (centros antes del PAE) y un escenario posprograma (centros después del PAE), y que procedemos a comparar los resultados escolares de los alumnos vulnerables en los dos escenarios. Imaginemos que este esquema —que denominamos antes-después— nos permite detectar un incremento significativo de estos resultados en el escenario posprograma; es decir, se produce una mejora en el rendimiento y niveles de graduación del alumnado vulnerable cuando los centros empiezan a desarrollar el PAE. ¿Es este cambio consecuencia del programa? La respuesta volvería a ser: «No necesariamente».

Las asociaciones entre resultados pre y posprograma, especialmente cuando lo único que se observa son los colegios afectados por el programa, vuelven a encontrarse sujetas a las limitaciones propias de la posible omisión de variables relevantes. Entre el antes y el después del programa, coincidiendo justamente con su lanzamiento e implementación, pueden producirse determinadas circunstancias o cambios en el contexto que, relacionados o no con el desarrollo del programa, sean en la práctica las verdaderas causantes de las mejoras en los resultados observados posprograma. Imaginemos que el PAE comenzó a implantarse en Cataluña en el año 2005, y que pocos años después (y aquí no hay que imaginarse nada), a partir de 2008, empieza a evidenciarse un descenso significativo del número de alumnos que abandonan prematuramente los estudios. La tentación de atribuir el mérito (o una parte del mérito) de este descenso al programa es grande. Y, sin embargo, la asociación antes-después, por sí sola, no nos permite determinar si el programa ha tenido mucho, poco o nada que ver.

De hecho, parecería razonable nombrar como principal responsable de este descenso la contracción que la crisis económica ha generado en la oferta de ocupación y que ha reducido drásticamente el coste de oportunidad de estudiar: lo que en el pasado representaba un incentivo a abandonar los estudios (facilidad para encontrar un empleo remunerado) se transforma ahora en un incentivo a permanecer en o volver al sistema escolar (dificultades para encontrar trabajo, especialmente para los jóvenes menos formados).

El ejercicio de comparar los resultados en cuestión antes y después de una determinada actuación educativa puede revestir un gran interés en el marco de una evaluación de impacto. No obstante, este ejercicio por sí solo no es suficiente para pronunciarse sobre la efectividad de la actuación.

2.4 ENTONCES, ¿DE QUÉ PODEMOS FIARNOS? EVIDENCIA EXPERIMENTAL Y CUASIEXPERIMENTAL

Para saber si un programa como el PAE funciona, es decir, si es efectivo, si produce impactos positivos relevantes, lo que realmente deberíamos hacer es comparar lo que les pasa a los alumnos o colegios que participan en él (escenario factual) con lo que les hubiera pasado si no hubieran participado (escenario contrafactual). Entonces sí que estaríamos comparando poblaciones idénticas; lo único que las diferenciaría es la presencia o ausencia de la intervención en cuestión. Por tanto, cualquier diferencia observada entre los resultados de los alumnos y los colegios en uno y otro escenario podría atribuirse causalmente al programa (Blasco & Casado, 2009).

Está claro que el escenario contrafactual no es directamente observable: los mismos individuos no pueden ser y no ser objeto de un mismo programa al mismo tiempo. Esto hace que el principal reto de toda evaluación de impacto que aspire a ser rigurosa consista en proponer estrategias que permitan identificar la mejor hipótesis contrafactual posible, lo que en la práctica quiere decir identificar un grupo de "control" o "comparación" lo más parecido posible al grupo que participa en el programa[5]. Cuando comparamos alumnos-PAE con alumnos-no-PAE, centros-PAE con centros-no-PAE, países-PAE con países-no-PAE, contextos escolares sin PAE que pasan a tener PAE, hay que garantizar que lo que distingue a las unidades tratadas de las unidades control es que las primeras han recibido el programa y las segundas no; en el resto de características, tanto observables como no observables, unas unidades y otras deberían ser lo más equivalentes posible. En el campo educativo —como en el ámbito de las ciencias sociales en general— esto no siempre es fácil de conseguir.

Y, sin embargo, son diversas las estrategias que permiten avanzar en la identificación de escenarios contrafactuales, de grupos de control, creíbles. La más robusta pasa por asignar aleatoriamente las unidades en disposición de acceder al programa entre grupo de tratamiento y grupo de control. Este procedimiento es la piedra angular del diseño experimental en evaluación, diseño que algunos autores han calificado como el gold standard a la hora de

aportar evidencias sobre qué funciona (véase, por ejemplo, Duflo, Glennerster, & Kremer, 2007). Cuando la aplicación de este diseño, por unos u otros motivos, no es viable, pueden entrar en juego determinados diseños de evaluación cuasiexperimental con contrafactual, en particular, aquellos diseños susceptibles de minimizar los sesgos de selección que hemos mencionado antes.

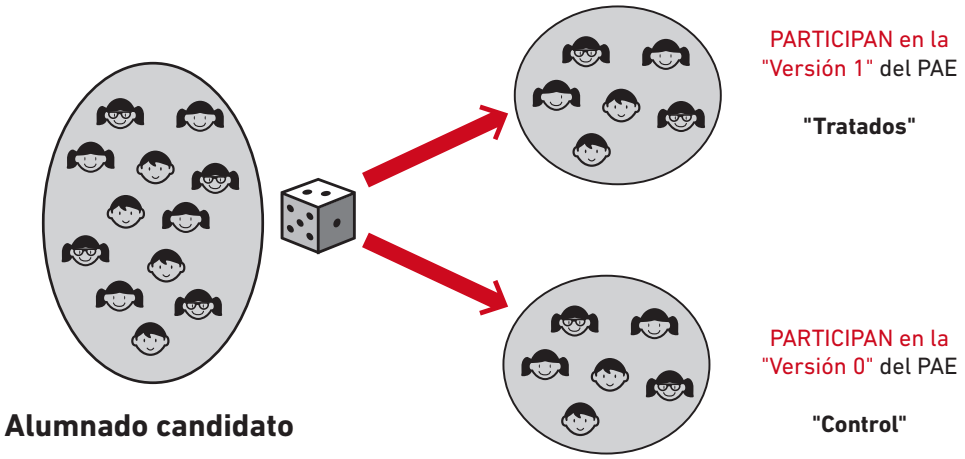
Las notas que siguen presentan algunos de los principales rasgos característicos de los distintos métodos de evaluación. Estas notas no quieren ni pueden ser exhaustivas. Su propósito es proporcionar elementos informativos que faciliten la interpretación de los ejemplos de evaluaciones expuestos en apartados posteriores y, al mismo tiempo, sus potenciales aplicaciones en Cataluña.

2.4.1 NOTAS SOBRE LA EVALUACIÓN EXPERIMENTAL

Evaluar el impacto de una política educativa mediante un diseño experimental es, desde una perspectiva metodológica, muy similar a aplicar la lógica que siguen los ensayos clínicos para probar la efectividad de un fármaco. La diferencia radica en que lo que evaluamos no son los efectos de un medicamento sobre el estado de salud, sino la capacidad de una intervención educativa para influir positivamente sobre una determinada problemática. En ambos casos, no obstante, el elemento central lo encontramos en la aleatorización de la participación, es decir, en la aplicación de un mecanismo de sorteo como procedimiento para asignar, de entre el conjunto de unidades elegibles para el tratamiento, cuáles acaban recibiéndolo y cuáles se incorporan al grupo de control.

En nuestro ejemplo, si pusiéramos la atención en la unidad «colegio», se trataría de escoger aleatoriamente, entre el conjunto de centros que han solicitado desarrollar el PAE y reúnen las condiciones formales requeridas por la convocatoria, cuáles se adhieren al programa (grupo de tratamiento) y cuáles no (y pasan a formar parte del grupo de control). El procedimiento sería el mismo en el caso de que la unidad de observación la pusiéramos en el alumno: centro por centro, escogeríamos de forma aleatoria qué alumnos, entre todos los que solicitan participar y reúnen los requisitos establecidos, reciben el programa y cuáles son asignados al grupo de control (véase la figura 1).

Figura 1. El diseño experimental. Programa frente a No programa



Fuente: *elaboración propia*

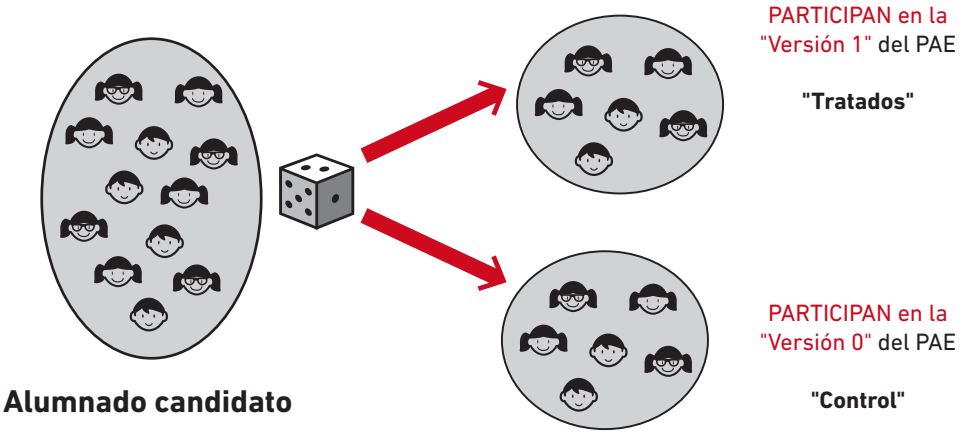
Cuando el número de casos es lo bastante relevante, la aleatorización garantiza que el grupo de individuos o unidades control y el grupo de tratamiento acaban siendo equivalentes en todas aquellas características, observables y no observables, que pueden estar relacionadas con los resultados considerados. Lo único que les diferenciará entonces será el hecho de haber participado o no en el programa, causa de toda posible diferencia en estos resultados. Más particularmente, el hecho de que la asignación aleatoria se realice entre solicitantes activos de un determinado recurso o servicio permite comparar conjuntos de individuos con un mismo nivel de motivación, al menos el nivel de motivación vinculado al acto de solicitarlo. El que la definición de las características que deben reunir los individuos solicitantes de la ayuda (población elegible) sea más o menos restringida no tiene por qué cuestionar la robustez de los resultados de la evaluación de sus impactos.

Hay que decir que el uso del diseño experimental en la evaluación de políticas públicas y sociales es objeto de críticas diversas. Dos de ellas han tenido una repercusión especial. La primera de ellas cuestiona el sustrato ético de la experimentación social: resulta injusto privar a determinados individuos (los del grupo de control) de los beneficios que supone una nueva política o servicio utilizando un mecanismo tan arbitrario como la aleatorización. El fundamento de este argumento es, sin embargo, criticable. En primer lugar, la presunción de que se está privando a algunos individuos de algo beneficioso pierde su sentido si pensamos que es precisamente la ausencia de evidencias sobre la efectividad del programa lo que justifica el experimento. En segundo lugar, en situación de escasez de recursos, cuando un determinado servicio o programa no puede cubrir el conjunto de la población a la que se dirige

(en el caso del PAE: conjunto de alumnos al inicio de la ESO en todos los institutos y colegios de Cataluña), entonces la asignación aleatoria del recurso entre la población que lo solicita y que es igualmente elegible es seguramente el mecanismo de distribución más justo que podemos diseñar.

En todo caso, cuando un programa se basa en evidencias prometedoras y cuando se dispone de recursos para dotarlo de una cobertura relativamente amplia, entonces pueden entrar en juego dos diseños experimentales seguramente más aceptables. Un primer diseño pasaría por aleatorizar la participación en distintas versiones del programa (véase la figura 2). De esta manera, no se mide el impacto de un programa respecto a una situación de ausencia de programa, sino hasta qué punto unas versiones del programa son más efectivas que otras.

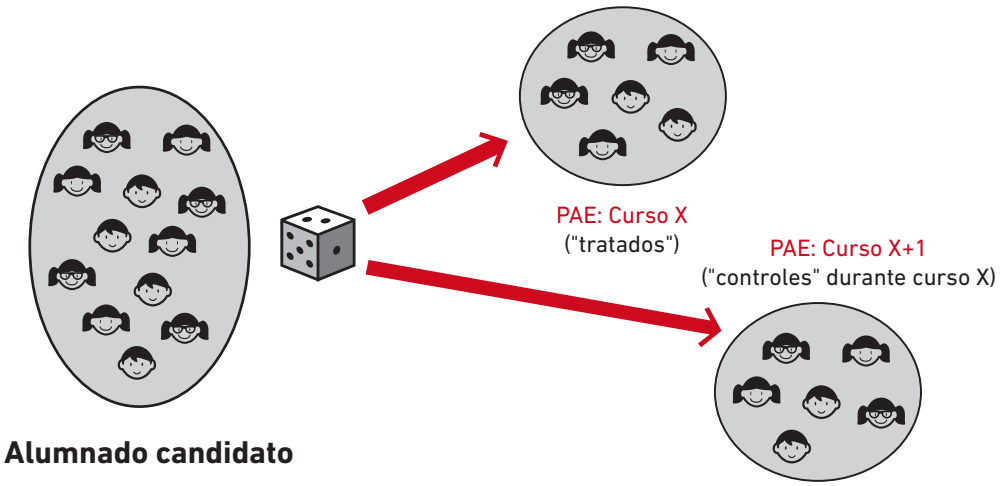
Figura 2. El diseño experimental. Programa 1 vs. Programa 0



Fuente: *elaboración propia*

Una segunda alternativa pasaría por aleatorizar el momento y orden de entrada en el programa (véase la figura 3). En este caso, un sorteo decide, entre todos los candidatos a participar en el programa, quién accede a él primero y quién después (al término de un trimestre, de un semestre o de un año académico). Los individuos que son asignados a la segunda edición del programa, mientras permanecen en espera, actúan como control de aquellos otros asignados a participar en él en primera instancia.

Figura 3. El diseño experimental. Programa en el momento X frente a Programa en el momento X + 1



Fuente: *elaboración propia*

La segunda crítica habitual a los experimentos sociales remite a la falta de validez externa de sus resultados, es decir, al hecho de que las conclusiones que se obtienen de ellos, si bien son válidas respecto a la población, lugar y momento implicados en el experimento (es decir, a pesar de tener validez interna), pueden no ser extrapolables a otros contextos distintos. Este es un argumento fuerte, que, sin embargo, puede contrapesarse en dos direcciones. En primer lugar, son habituales los denominados pilotos multi-site en los que un mismo programa se evalúa en distintos lugares (varios municipios, por ejemplo) con el propósito de analizar hasta qué punto los resultados de impacto varían en función de los contextos. En segundo lugar, cuando el número de réplicas experimentales de un determinado tipo de intervención es lo suficientemente importante, puede realizarse lo que se denomina metanálisis de los resultados obtenidos, es decir, un ejercicio cuantitativo de síntesis que pretende establecer si programas del tipo considerado resultan efectivos con carácter general, al margen de las poblaciones, lugares y momentos en que se apliquen.

Resulta evidente que las evaluaciones experimentales plantean retos importantes tanto de carácter metodológico como de viabilidad técnica. Sin embargo, lo sorprendente es que un diseño evaluativo como este, cada vez más utilizado en otros países y considerado a nivel internacional el método más robusto de la evaluación de impacto se encuentre prácticamente ausente de la evaluación de políticas públicas en Cataluña (Casado, 2012)[6].

2.4.2 NOTA SOBRE LOS DISEÑOS CUASIEXPERIMENTALES

Los métodos cuasiexperimentales de evaluación de impacto comparten con el diseño experimental la definición de un grupo de control o comparación (no participantes en el programa) de cara a estimar el outcome contrafactual de los participantes. No obstante, a diferencia del diseño experimental que elimina el sesgo de selección mediante la aleatorización de la participación, el resto de métodos solo logra este objetivo si se cumplen ciertos supuestos en lo que respecta al mecanismo que gobierna la participación en el programa.

Son diversos los métodos cuasiexperimentales que se utilizan para evaluar la efectividad de programas e intervenciones educativas. Tres de ellos, la regresión discontinua, los modelos de dobles diferencias y los efectos fijos, disponen de aplicaciones interesantes en este ámbito.

Regresión discontínua

Es probable que esta sea la técnica de evaluación no experimental que más proyección ha tenido en el ámbito educativo en los últimos años (Schlotter, Schwerdt, & Woessmann, 2010; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). La regresión discontinua puede aplicarse cuando la participación en el programa en cuestión depende del hecho de que una variable tome un conjunto de valores determinados, es decir, cuando la posibilidad de acceder al programa se establece a partir de un corte en una variable numérica. Más concretamente, esta debería ser una variable presumiblemente vinculada a los outcomes de interés.

Por ejemplo, podría darse el caso de que, para avalar la existencia de consenso en la decisión de los centros de solicitar el PAE, se requiriera un mínimo de un 75 % de los votos del consejo escolar a favor de esta decisión. O podría establecerse que fueran elegibles para recibir el programa únicamente los institutos con, como mínimo, un 25 % de alumnos al inicio de la ESO con carencias acreditadas en las materias instrumentales. Más aún, la administración podría recomendar a los centros PAE que priorizan la participación en el programa de aquellos estudiantes que hubieran puntuado por debajo de un determinado umbral en las pruebas de competencias básicas de 6.º de primaria.

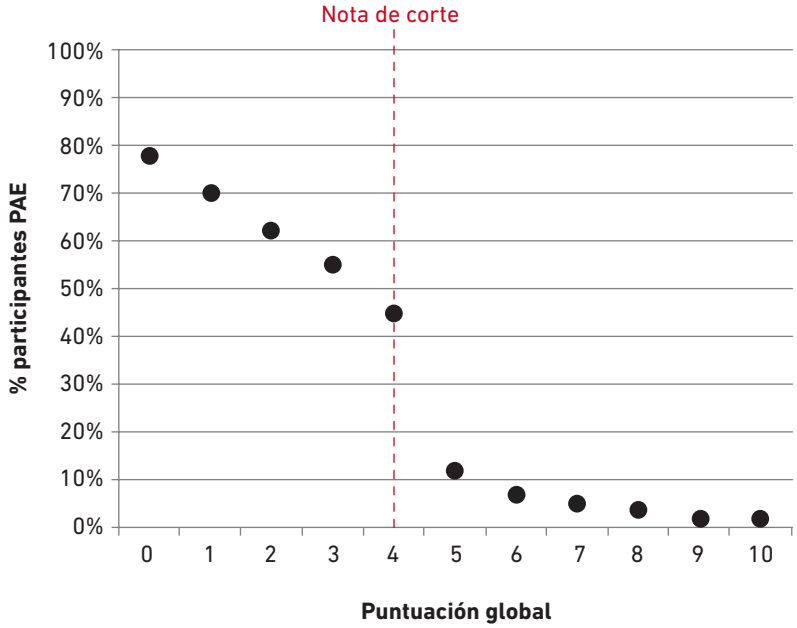
En tanto que la definición de estos umbrales ha sido establecida de forma exógena (por determinación de la administración educativa), podríamos esperar que los centros y los alumnos que se encuentren justo por debajo y justo por encima del punto de corte serán esencialmente equivalentes en todos los atributos relevantes, tanto observables como inobservables.

Por ejemplo, imaginemos que el Departament d'Ensenyament adjudica un número limitado de plazas PAE a cada uno de los centros implicados y les pide que prioricen la atención de aquellos alumnos que obtuvieron menos de cuatro puntos de calificación global (en una escala de 0 a

10) en las pruebas de competencias básicas de 6.º de primaria; recordemos que todos ellos son alumnos que han solicitado acceder al programa.

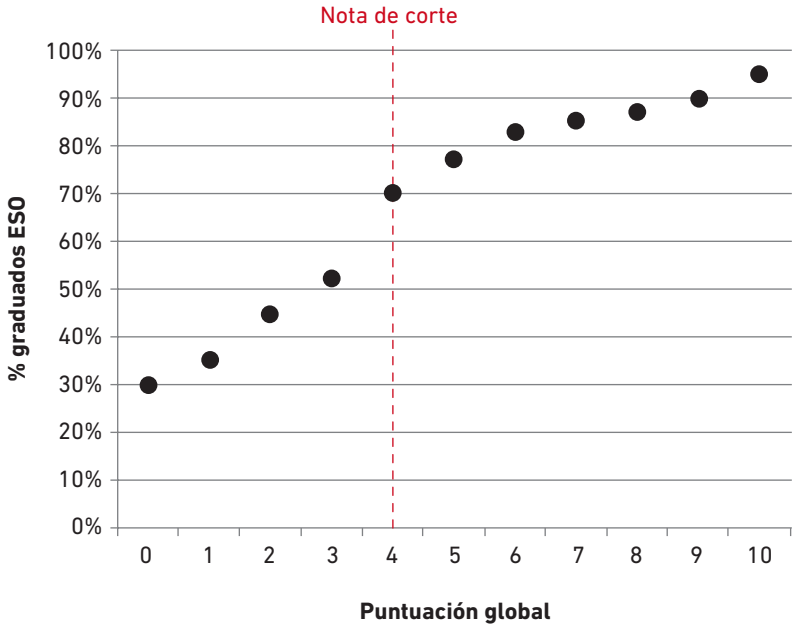
Deberíamos entonces constatar que el porcentaje de alumnos PAE da un salto cuando estos superan este umbral de calificación (gráfico 1). De esta forma, si el programa es efectivo, deberíamos poder observar que el salto en el porcentaje de participación se corresponde con otro salto en el nivel de rendimiento de los alumnos con puntuaciones de partida en torno a este umbral (gráfico 2). Este segundo salto reflejaría el impacto del programa.

Gráfico 1. Proporción de alumnos PAE según puntuación global en pruebas de competencias en 6.º de primaria



Funte: *elaboración propia (datos no reales)*

Gráfico 2. Proporción de alumnos que obtienen el GESO según puntuación global en pruebas de competencias en 6.º de primaria



Fuente: *elaboración propia (datos no reales)*

El método de regresión discontinua proporciona buenas estimaciones del impacto de los programas sobre la población que se encuentra cerca del umbral de referencia; no en vano, es la proximidad al umbral lo que garantiza la similitud entre participantes y no participantes[7].

Variables instrumentales

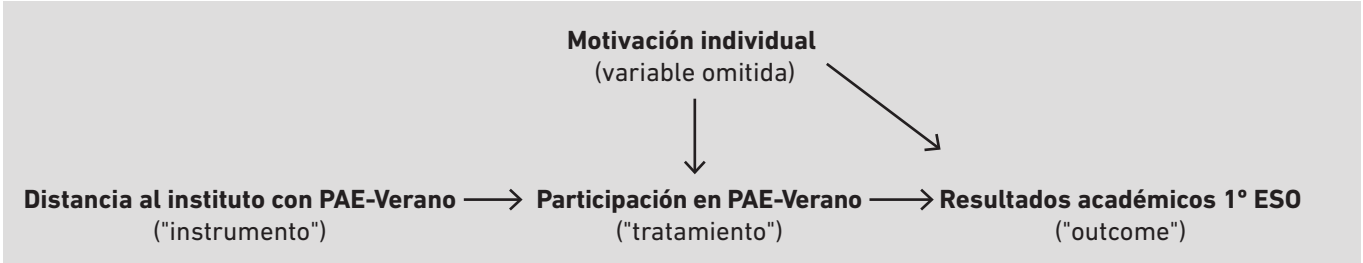
En el marco de la evaluación de impactos, una variable instrumental (o instrumento) es un factor que está claramente vinculado a la probabilidad de participar en un programa, pero que no mantiene ninguna asociación directa con los indicadores de outcome que se consideran; en otras palabras, el instrumento es algo exógeno al modelo explicativo en cuestión (Imbens, 2014; Porter, 2012).

Son diversas las variables instrumentales que han sido utilizadas para evaluar el impacto de políticas y programas educativos (Schlotter et al., 2010). Algunos estudios han considerado cambios legislativos o normativos que modifican la probabilidad de acceder a un determinado programa o de verse afectado por una determinada política (cambios sin relación directa con los outcomes de interés); por ejemplo, un cambio en la duración de la etapa de escolarización obligatoria, la cual provoca variaciones en los años de escolarización de los individuos de distintas cohortes que van más allá de sus habilidades o disposiciones escolares.

En otros contextos, se han instrumentado factores «naturales» o de tipo geográfico. Este es, por ejemplo, el caso de la distancia geográfica entre la residencia de los individuos y el lugar donde se ofrece el programa. La hipótesis de partida es que, a igual distancia entre residencia y localización del programa en cuestión, el hecho de que unos individuos se dirijan a él y otros no, no es aleatorio. Sobre todo, cuando estamos ante programas voluntarios, este hecho nos habla de diferencias entre los atributos de los distintos individuos que resultan a menudo difíciles de objetivar (motivación, disponibilidad, iniciativa, etc.) y que, sin embargo, están vinculados a los outcomes en cuestión. En cambio, asumimos que a iguales características (observables y no observables), los individuos que viven cerca del lugar (o lugares) donde se ofrece el programa tendrán una probabilidad más alta de participar en él que los que viven lejos. No en vano, la distancia geográfica puede implicar claras dificultades objetivas de acceso al programa.

Supongamos que el Departament d'Ensenyament decide plantear una ampliación del PAE al período de vacaciones de verano, centrándose en la atención a los alumnos que comenzarán la ESO en septiembre y que tienen un nivel bajo en competencias instrumentales. El refuerzo escolar lo realizarán estudiantes universitarios voluntarios a razón de dos horas diarias a lo largo del mes de julio. Imaginemos también que, por un motivo u otro, el Departament decide ofrecer el programa solamente en algunos institutos de cada municipio, abriendo el acceso, eso sí, a cualquier alumno residente en el municipio con independencia del centro de secundaria que tenga asignado. En este caso, la distancia entre el lugar de residencia de los alumnos y los institutos con PAE-Verano podría llegar a condicionar el mero hecho de interesarse por él. Por el contrario, si la elección del instituto en cuestión satisface ciertas condiciones[8], sería esperable que el hecho de vivir más o menos lejos de este instituto, a igualdad de otras características sociodemográficas, no mantenga ninguna relación significativa con los futuros resultados académicos de los alumnos (figura 4). Así, con la ayuda de este instrumento inferiríamos el impacto del programa controlando por aquella probabilidad de participar en él que depende de variaciones en la distancia entre residencia e instituto con PAE-Verano.

Figura 4. La evaluación del PAE, con la distancia como variables instrumental: diseño general



Fuente: *elaboración propia*

La robustez de las estimaciones de impacto basadas en el uso de variables instrumentales dependerá de la credibilidad del instrumento en sí, es decir, de la medida en que este sea capaz de lograr la doble condición de vinculación con la participación y no vinculación con los outcomes.

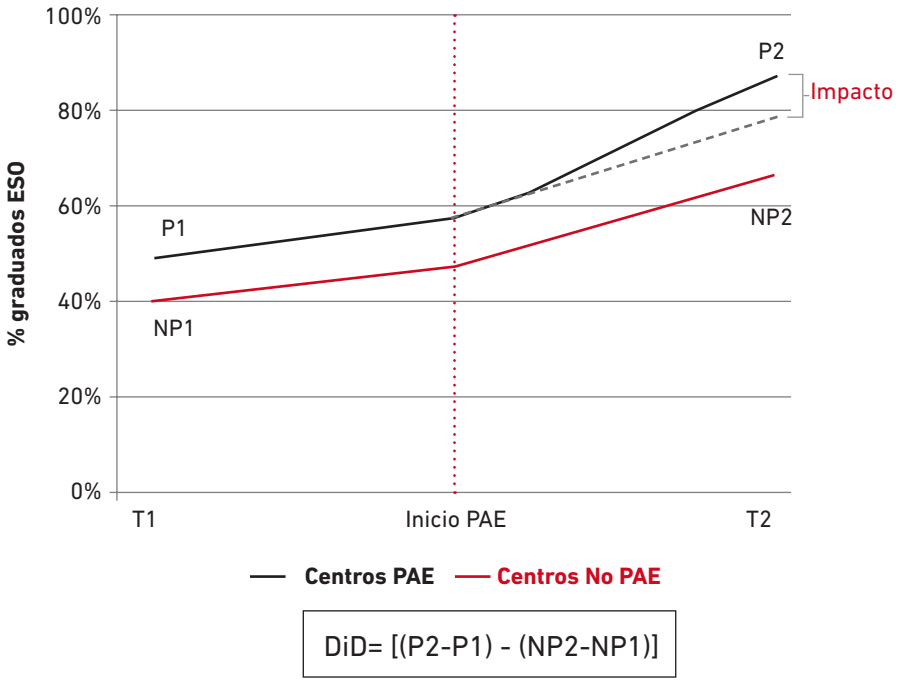
Dobles diferencias

Los modelos de dobles diferencias representan un paso más allá del diseño "antes-después". A diferencia de este diseño, la técnica de dobles diferencias (o diferencias-en-diferencias) trata de controlar los efectos de los factores contemporáneos al programa comparando lo que les sucede a participantes y no participantes a lo largo del mismo período de tiempo.

La aplicación más habitual de los modelos de dobles diferencias es la que establece la comparación entre dos grupos de individuos (o de colegios, regiones, países...) a lo largo del tiempo: en el primer período, ninguno de los dos grupos participa en el programa; en el segundo período, en cambio, un grupo participa en el (tratamiento) y el otro no (comparación). La estrategia de estimación del impacto que proponen estos modelos se basa en la comparación de las diferencias en los outcomes de cada grupo antes y después de la participación en el programa. Asumiendo que el sesgo de selección entre ambos grupos permanece constante a lo largo del tiempo, la diferencia entre estas dos diferencias señalaría el impacto del programa.

Aplicado a nuestro ejemplo (véase el gráfico 3), y fijado al nivel del colegio, la técnica de las dobles diferencias consistiría en: 1) calcular la variación entre los resultados de los alumnos PAE antes y después de comenzar el programa, así como la variación entre los resultados de alumnos de iguales características (curso, nivel académico, extracción socioeconómica, etc.) en centros no-PAE también antes y después de la puesta en marcha del programa; 2) calcular la diferencia entre estas dos variaciones. De nuevo, supondremos que cualquier diferencia previa al programa entre unos centros y otros, ya sea grande o pequeña, se habría mantenido constante aunque el programa no hubiera existido.

Gráfico 3. Evolución de la proporción de alumnos-PAE y no-PAE que obtienen el GESO



Fuente: *elaboración propia (datos no reales)*

El principal punto débil de los modelos de dobles diferencias es la posible existencia de factores inobservables que varíen a lo largo del tiempo y que lo hagan con intensidades distintas entre participantes y no participantes. Así pues, si la motivación de alumnos-PAE y no-PAE varía a lo largo del tiempo, y no podemos garantizar que lo haga en la misma medida entre unos y otros, no podremos estar seguros de que las diferencias en sus resultados después del programa no tengan su origen en estas variaciones. Si queremos que resulten creíbles los resultados de una evaluación de impacto que utilice un diseño de dobles diferencias, habrá que ofrecer argumentos que permitan descartar la existencia de esta limitación.

Tabla 1. Cuatro métodos de evaluación de impacto: características básicas

MÉTODO	*RASGOS PRINCIPALES*	*VENTAJAS*	*REQUERIMIENTOS*	*LIMITACIONES*
Aleatorización – diseño experimental	Requiere la asignación aleatoria de la población diana al grupo de tratamiento y al grupo de control	• Si se implementa correctamente, proporciona las estimaciones de impacto más robustas, ya que elimina de raíz el sesgo de selección • Los resultados que proporciona son transparentes, fáciles de entender y difíciles de cuestionar • Existen diferentes diseños que se pueden adaptar al contexto de la política y las circunstancias de la intervención	• Requisitos básicos para controlar la asignación aleatoria a la participación • Registro estricto de las asignaciones a cada uno de los grupos • Conveniencia de recoger datos basales • Se requieren medidas de resultado para el grupo de tratamiento y de control	• Rechazo a formar parte del grupo de control • Consentimiento de los participantes no siempre fácil de obtener • La aleatorización puede influir en el perfil de los candidatos que se postulan al experimento • El hecho de que los participantes conozcan su estatus como tratamiento o control puede alterar su comportamiento e influir en los resultados • Reservas éticas a la aleatorización • Volumen importante de requisitos de diseño i planificación
Diseños de regresión discontinua	Los individuos reciben la intervención según si su puntuación en una variable continua se encuentra por encima o por debajo de un umbral predeterminado. Este punto de corte marca la distinción entre grupo de tratamiento y comparación	• Se puede aplicar tanto cuando la discontinuidad en torno al punto de corte es nítida y dicotómica (diseño "sharp") como cuando es difusa (diseño "fuzzy") • Bajo ciertas condiciones, proporciona estimaciones de impacto robustas (no sesgadas)	• La determinación del punto de corte debe ser independiente de los valores que se asignan a los individuos de la población diana • Se requieren datos sobre los valores en la variable de asignación y medidas de resultado para los individuos del grupo de tratamiento y comparación	• El método no es viable sin la existencia de una variable continua de asignación • Los análisis pueden resultar complejos y en ocasiones inciertos cuando las razones que gobiernan el establecimiento del punto de corte no son claras o cuando la muestra de individuos alrededor del umbral es reducida • Pueden plantearse dificultades a la hora de interpretar los resultados de las estimaciones y de generalizar las conclusiones extraídas

MÉTODO	*RASGOS PRINCIPALES*	*VENTAJAS*	*REQUERIMIENTOS*	*LIMITACIONES*
Variables instrumentales	Se utiliza una variable o instrumento como fuente de variación en la participación en el programa. El instrumento debe ser exógeno, esto es, no relacionado con los resultados considerados	• Bajo ciertas condiciones, proporciona estimaciones de impacto robustas (idea de experimento natural), porque resuelve el problema del sesgo de selección • Se puede aplicar en avaluaciones retrospectivas • Permite la estimación de diferentes tipos de impacto	• Requiere datos basales, de resultado y valores relativos a la variable instrumental para los individuos del grupo de tratamiento y comparación • El instrumento debe tener relación con la probabilidad de recibir el tratamiento y afectar a los resultados sólo a través de la determinación de esta probabilidad • El instrumento no puede asociarse a ningún otro determinante de los resultados	• A menudo es difícil llegar a identificar variables instrumentales válidas • No es fácil explicar el método a los no expertos • La interpretación de los resultados no es fácil ni directa • La posibilidad de testar los supuestos que justifican la selección del instrumento es a menudo limitada
Modelos dobles diferencias	Se comparan las medidas de resultado de participantes y no participantes antes y después de la intervención	• Se controlan determinados factores inobservables que pueden generar diferencias sistemáticas entre participantes y no participantes • Se puede utilizar en conjunción con técnicas de emparejamiento ("matching") • Se puede utilizar con datos de sección cruzada ("cross-section") o de panel	• Requiere datos basales y de resultado de participantes y no participante antes y después de la intervención • Para validar el método, se pueden requerir múltiples observaciones pre-intervención, tanto para participantes como para no participantes	• Se asume que los resultados de participantes y participantes siguen patrones de evolución comunes • Los análisis acostumbran a ser complejos y no siempre fáciles de comunicar • No es siempre fácil obtener la diversidad de datos pre-intervención requeridos para validar el método • No se puede utilizar para estimar efectos múltiples de la intervención

Fuente: *adaptación de Morris, Tödling-Schönhofer, & Wiseman, 2013, pp. 24-25.*

Notas:

[2] *El parecido entre el hipotético programa PAE y el existente PIM (Programa Intensivo de Mejora, abierto a los institutos catalanes de secundaria desde el curso 2012-2013) no es casual.*

[3] *De forma general, existe endogeneidad cuando la variable de tratamiento (recibir o no recibir el PAE) no es independiente (exógena) del outcome considerado y del contexto que lo condiciona. Las fuentes posibles de endogeneidad son diversas. En el campo que nos ocupa, es habitual encontrarse con dos de ellas: cuando la variable de tratamiento se encuentra determinada por el outcome en sí (endogeneidad por causalidad inversa) y cuando el tratamiento tiene relación con variables no observadas igualmente relacionadas con el outcome (endogeneidad por variables omitidas) (Schlotter, Schwerdt, & Woessmann, 2010).*

[4] *El sesgo de selección es un caso concreto de endogeneidad generada por variables omitidas, habitual en el marco de la evaluación de políticas y programas. El sesgo de selección se produce cuando existen diferencias entre tratados y no tratados que son previas al inicio del programa y que pueden ser responsables de las diferencias entre los resultados de unos y otros cuando este finaliza.*

[5] *El uso del término grupo de control suele reservarse al marco de la evaluación experimental; la expresión grupo de comparación es más propia de los estudios no experimentales.*

[6] *A fecha de marzo de 2015, y hasta donde llega nuestro conocimiento, los únicos experimentos en marcha en Cataluña para evaluar el impacto de programas educativos los encontramos en la evaluación del programa Mobilitza't-Mobile (Barcelona Activa) y la Maleta de les Famílies (Diputación de Barcelona). Mobilitza't es un programa de formación e inserción laboral en el sector del móvil dirigido a jóvenes con un bajo nivel de estudios; el programa arrancó a finales de 2013 y se desarrollará hasta finales de 2016. El programa La Maleta de les Famílies (La Maleta de las Familias) consiste en la programación de talleres grupales, tutorías y seguimientos individuales para familias con hijos en 4.º de ESO en riesgo de abandono educativo; su implementación se realiza durante los cursos 2014-2015 y 2015-2016. Ambos programas se plantean en clave de pilotos experimentales y están siendo evaluados por Ivàlua.*

[7] *La validez del método requiere la satisfacción de dos supuestos básicos: que ninguna otra variable relevante salte en el mismo punto de corte y que los individuos no escojan libremente en qué lado del punto de corte se ubican. Por todo ello, estudios basados en la regresión discontinua son a menudo calificados de experimentos naturales.*

[8] *Por ejemplo, que no haya venido guiada por la presión de colectivos de padres y madres especialmente motivados en los distintos barrios.*

3. LOS RETOS DE LA EXPERIMENTACIÓN: EJEMPLOS

A nivel internacional, existe actualmente un repertorio bastante extenso de experimentos llevados a cabo en distintos ámbitos de la política educativa. En el mundo occidental, buena parte de estos experimentos han sido desarrollados en los Estados Unidos y, más recientemente, en el Reino Unido. Aun así, muchos de ellos permiten ilustrar cuáles podrían ser los retos y las potencialidades del empleo de este método en la evaluación de políticas educativas en Cataluña. La selección de ejemplos que describiremos a continuación persigue justamente este propósito.

3.1 LA AUTONOMÍA ESCOLAR: LOS EXPERIMENTOS DE LAS CHARTER SCHOOLS

La delegación de más autonomía a los centros escolares ha pasado a ocupar una posición central en el debate educativo a nivel global. Pero, ¿de qué tipo de autonomía estamos hablando? ¿Y qué nos dice la evidencia internacional sobre la efectividad de unas formas de autonomía u otras?

Un buen número de estudios basados en pruebas internacionales como las de PISA, TIMSS o PIRLS[9] llegan a la conclusión de que el rendimiento medio de los estudiantes mejora en aquellos centros que disponen de un amplio margen de autonomía en la administración de los recursos (en particular, en la asignación del presupuesto disponible), en la gestión del profesorado (qué profesores contratar y cómo incentivarlos) y en la selección de métodos pedagógicos (Woessmann, 2003; Fuchs & Woessmann, 2007; Woessmann, Luedemann, & Schuetz, 2009). En cambio, se observa que no siempre lo que puede ser positivo en términos de rendimiento medio lo es también en el terreno de la equidad educativa (Schütz, West, & Woessmann, 2007). Así pues, el peso del origen social en la explicación de los logros escolares se incrementa en los países donde una parte importante de los colegios puede interceder en el proceso de selección y admisión de los alumnos (y, por tanto, caer en la tentación del cream-skimming[10]). Otros estudios evidencian una clara interacción entre autonomía escolar y determinados mecanismos de rendición de cuentas (accountability), en el sentido de que la autonomía funciona mejor —en términos de nivel medio y de equidad de resultados— siempre y cuando los colegios sean sometidos a estos mecanismos (Woessmann, 2005b). Finalmente, algunos autores han constatado que los efectos positivos que tiene asociados la autonomía escolar en los países ricos desaparecen en entornos empobrecidos (Hanushek, Link, & Woessmann, 2013).

En cualquier caso, deberíamos recordar que estos estudios difícilmente consiguen neutralizar los problemas de endogeneidad propios de la comparativa internacional mediante datos de carácter transversal, de tal modo que difícilmente llegan a identificar relaciones de causalidad entre política y resultados de interés (outcomes).

Una manera de eludir estas limitaciones pasa por aprovechar la variabilidad que este factor presenta a menudo dentro de algunos países o regiones. Particularmente interesante es la posibilidad de sacar partido al hecho de que algunos centros (habitualmente los que disponen de un mayor nivel de autonomía) disponen de sistemas de admisión específicos, incorporando a menudo un mecanismo de sorteo. Un caso paradigmático de esta situación es el de las charter schools.

Los colegios charter empezaron a implantarse en los Estados Unidos a principios de los años 1990. Aunque son centros de iniciativa y gestión fundamentalmente privadas (sin ánimo de lucro), son definidos como parte integrante del sistema público de educación. Gran parte de su financiación procede de fondos públicos, a cambio de que garanticen la gratuidad en el acceso. Las charter disponen de un nivel de autonomía escolar (profesorado, presupuesto, currículo y pedagogía) muy superior al de los colegios públicos convencionales, y su elección no está sujeta a ninguna restricción de asignación zonal; en caso de sobredemanda, el acceso se determina mediante un sorteo. Esta última circunstancia brinda una clara oportunidad al diseño de experimentos dirigidos a evaluar los impactos de estos centros (colegios con un estatus especial de autonomía) sobre distintos outcomes de interés.

Un balance de la evidencia acumulada a día de hoy señalaría que la efectividad de este tipo de colegio acostumbra a ser muy variable. Se constata que las charter funcionan en determinados contextos, pero en otros no, funcionan para determinados grupos de alumnos, pero no para otros (Betts & Tang, 2011; Di Carlo, 2011; CREDO, 2009; Zimmer et al., 2009).

Fijémonos, por ejemplo, en el estudio de Clark et al. (2014) sobre el impacto de 33 colegios charter de secundaria. Estos centros, todos ellos significativamente sobredemandados, se distribuyen en trece estados distintos de los EE. UU. (urbanos y rurales) y presentan características sociodemográficas diversas. Como en todos los experimentos sobre charter schools, los autores utilizan los resultados del sorteo de acceso a estos colegios para identificar al grupo de control correspondiente, los «perdedores» de la lotería. Este mecanismo de aleatorización garantiza que este grupo es equivalente al grupo de tratamiento (los «ganadores» de la lotería) en todas aquellas características inobservables que pueden tener que ver con la voluntad de escolarizarse en una charter. En conjunto, el estudio identifica 1400 alumnos tratados y 930 controles, y sigue la evolución de sus logros escolares a lo largo de dos cursos. Como resultado general, se constata que, de media, las charter analizadas no generan ningún impacto significativo (ni positivo ni negativo) en el rendimiento académico global de los alumnos que se escolarizan en ellos.

En cambio, estudios como los de Angrist et al. (2013) y Abdulkadiroglu et al. (2009) sobre los colegios charter de Boston, o como los de Hoxby et al. (2009) y Dobbie y Fryer (2011) sobre las charter de la ciudad de Nueva York, todos basados en la misma estrategia de identificación (ganadores y perdedores de la lotería de acceso), aportan evidencias sólidas de casos de éxito

de este perfil de colegios, principalmente en la mejora de los resultados académicos y de continuidad educativa de los alumnos más desfavorecidos[11].

Hay que decir que esta variabilidad en la efectividad de las charter schools ha conducido a no pocos expertos a sostener el argumento de que, charter o no charter, lo que importa es lo que el colegio hace con su autonomía (Fryer, 2011a).

EN CATALUÑA...

- En Cataluña, en el año 2005, se puso en marcha el Proyecto para la Mejora de la Calidad de los Centros Educativos (Projecte per a la Millora de la Qualitat dels Centres Educatius), conocido como Proyecto de Autonomía de Centros (Projecte d'Autonomia de Centres, PAC). Este proyecto comportaba para los centros que se adherían a él una entrada importante de recursos (principalmente en forma de subsidio económico) y la delegación de un alto grado de autonomía en funciones relacionadas con los ámbitos pedagógico, organizativo y de gestión de recursos humanos y materiales. A cambio, los centros se abrían a la evaluación externa y a la rendición de cuentas. Entre 2005 y 2009, 635 centros escolares de primaria y secundaria participaron en el PAC. En el año 2009, la Ley de Educación de Cataluña (de 10 de julio de 2009) se encargaría de consagrar la autonomía escolar como uno de los principios rectores del sistema educativo, y el Decreto de Autonomía de los Centros Educativos (de 3 de agosto de 2010) de trasladar los criterios del PAC al conjunto de centros públicos y concertados de Cataluña.
- En este marco, se han logrado diversos avances en la delegación a los centros de un mayor grado de autonomía en sus distintas dimensiones; en la dimensión pedagógica mediante el fomento de los proyectos de innovación pedagógica (entre otros instrumentos); en las dimensiones organizativa y de gestión mediante el refuerzo y la profesionalización de las direcciones de los centros (entre otras medidas).
- No podemos decir, sin embargo, que la introducción de estas reformas haya venido acompañada de pruebas empíricas sólidas sobre su efectividad, como tampoco se han acabado de establecer las condiciones de diseño que deberían hacer posible su evaluación de impacto. Es especialmente destacable la falta de una evaluación rigurosa del PAC, sobre todo considerando que el plan fue formulado inicialmente como una prueba piloto.

3.2 EDUCACIÓN EN LA PRIMERA INFANCIA: PERRY PRESCHOOL Y HEAD START

Distintos estudios se han aproximado a los efectos de la educación preescolar —entendida como etapa no obligatoria de la educación infantil— sobre el progreso educativo de los alumnos comparando las realidades institucionales y los resultados educativos de distintos países y regiones (Schütz, 2009; Schütz, Ursprung, & Woessmann, 2008). Recordemos, no obstante, las limitaciones que presenta la comparación internacional como método para estimar impactos; limitaciones que se derivan de la potencial omisión de variables relevantes para el análisis. Por otro lado, en el nivel individual, hay que contar con la presencia de un sesgo de selección importante entre los colectivos que escolarizan a sus hijos a edades muy tempranas. Aunque consigamos comparar familias «usuarias» y «no usuarias» similares en lo observable (por ejemplo, características sociodemográficas), nada nos asegura que unas y otras no sean distintas en otros rasgos no observables que podrían estar relacionados con el outcome de interés (por ejemplo, respecto al valor que otorgan a la institución escolar o respecto a la capacidad de adoptar determinadas estrategias educativas en el ámbito doméstico).

Superando estas limitaciones, un buen número de estudios ha sacado partido del diseño experimental que ha guiado las pruebas piloto o el desarrollo de distintos programas de educación infantil en los Estados Unidos. Ejemplos clásicos de estas iniciativas son el Perry Preschool Program (iniciado el año 1961 en la ciudad de Ypsilanti, Michigan) y el Abecedarian Project (iniciado en 1972 en Carolina del Norte). Más reciente en el tiempo, situaríamos el proyecto de estudio experimental del Head Start (iniciado en 2002 a nivel federal). Nos referiremos brevemente al primero y al último de estos experimentos.

El experimento del Perry Preschool se articuló de la forma siguiente: 123 niños de tres y cuatro años procedentes de familias desfavorecidas fueron distribuidos aleatoriamente entre el grupo de tratamiento (58 niños), que incluye: programa de escolarización diaria y en grupos reducidos durante al menos un curso escolar, visitas de especialistas a domicilio y reuniones familiares colectivas periódicas; y el grupo de control (65 niños), que no recibirían ninguno de los servicios anteriores.

Distintos estudios han podido comparar la evolución de los resultados de ambos grupos a corto, medio y largo plazo (hasta llegar a los 40 años de edad de los participantes). De forma general, estas evaluaciones concluyen que el programa tuvo impactos positivos destacables en el ámbito de las capacidades cognitivas. Estos impactos, no obstante, se producían básicamente a corto plazo, y desaparecían durante los primeros cursos de la educación primaria. Por el contrario, se han podido documentar efectos del programa a largo plazo, en concreto, beneficios en el terreno laboral (más ingresos y más estabilidad) y en el social (menos prácticas de riesgo y delictivas) una vez en la edad adulta (Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010a, 2010b). Todo esto en conjunto explica los buenos resultados que el programa ha obtenido siempre que ha sido sometido a exámenes coste-beneficio (Barnett, 1985; Belfield, Nores, Barnett, & Schweinhart, 2006).

Las conclusiones a las que llegan las evaluaciones de impacto del programa Head Start coinciden, en parte, con los resultados del proyecto Perry School. Head Start es el programa de provisión de servicios de educación infantil de más alcance a nivel federal. Iniciado en el año 1965, en 2002 se puso en marcha un desarrollo experimental del programa, conocido como Head Start Impact Study. El experimento incluye a 23 estados, 383 centros preescolares y 4667 niños de tres y cuatro años de edad. Entre las familias interesadas se asigna aleatoriamente la posibilidad de inscribir a un hijo en un centro Head Start; a la mitad de ellas se les ofrece esta opción (niños tratados), al resto no (niños controles). Las últimas evaluaciones de este experimento concluyen (Peck & Bell, 2014; Puma et al., 2012): a) que el programa tiene impactos cognitivos positivos, sobre todo en el ámbito lingüístico, mientras dura la participación en el programa y hasta el inicio del primer curso de primaria; b) que los efectos no cognitivos son poco relevantes a corto plazo, pero que, en cambio, algunos de ellos (sobre todo los que se refieren el ámbito socioemocional) se hacen patentes en el momento de cursar el tercer año de primaria. En todo caso, cuando se detectan impactos (cognitivos o socioemocionales), son

especialmente remarcables entre los niños de familias socioeconómicamente más vulnerables. Evaluaciones cuasiexperimentales de ediciones anteriores del Head Start indican que las cohortes participantes vieron sensiblemente incrementadas sus probabilidades de llegar a la universidad y disminuidas sus tasas de criminalidad.

Qué hace que la escolarización en la primera infancia produzca beneficios cognitivos que desaparecen a corto plazo y, en cambio, muestra impactos positivos a largo plazo en distintas variables clave del ciclo vital (niveles de graduación, condiciones laborales, criminalidad, salud, etc.) es una cuestión a la que todavía no se ha encontrado respuesta (Duncan & Magnuson, 2013)[12]. En cambio, la literatura sí parece coincidir en la constatación de que son los niños de las familias socioeconómica y culturalmente más vulnerables los que más suelen beneficiarse de la escolarización a edades muy tempranas. Las principales revisiones sistemáticas sobre la cuestión (Camilli, Vargas, Ryan, & Barnett, 2010; Gorey, 2001; Manning, Homel, & Smith, 2010; Nelson, Westhues, & MacLeod, 2003) añaden dos puntos interesantes. Primero, que el alcance de los impactos de la educación infantil se incrementa cuando sigue determinados parámetros de calidad, por ejemplo, cuando se cuenta con un profesorado especialista y se trabaja en grupos reducidos; segundo, que la efectividad de esta educación aumenta a medida que se amplía el tiempo de exposición a este tipo de escolarización (horas al día y años de duración/edad de inicio).

EN CATALUÑA...

- Entre los años 2004 y 2008, el Gobierno catalán impulsó la ampliación del número de plazas de guardería pública (escola bressol) en el marco de lo que desde entonces se denomina el mapa de guarderías de Cataluña. Finalizado este período, el número de niños matriculados se había incrementado en 21.347. Desde entonces, y hasta el curso 2011-2012, la cobertura pública de plazas de guardería no paró de aumentar. A partir de entonces, esta cobertura inició un descenso significativo (Blasco, 2015).
- Asociado a este descenso en la cobertura encontramos las sucesivas reducciones de las aportaciones de la Generalitat a la financiación de las plazas públicas de guarderías. En tres cursos escolares (del 2010-2011 al 2012-2013), la Generalitat ha pasado de financiar el 50 % del coste de la plaza pública a la escolarización de 0 a 3 años a financiar el 25 %, dejando en manos de los municipios y las diputaciones la posibilidad de compensar esta reducción.
- Ni las medidas de impulso de la red de guarderías ni las acciones de contención del gasto en esta etapa educativa han sido planteadas apelando a evidencias empíricas sobre la mayor o menor trascendencia que este recurso puede tener en Cataluña. En concreto, conocer los impactos que esta escolarización puede tener a corto y medio plazo sobre la evolución educativa de niños y jóvenes debería contribuir a informar todo el análisis económico que pudiera llegarse a plantear en términos de relación coste-beneficio de los recursos destinados a esta intervención. Y el caso es que la efectividad de pocas políticas ha sido tan ampliamente avalada por la investigación como la de aquellas intervenciones dirigidas a ampliar la cobertura educativa en la primera infancia.

3.3 LO QUE SE PAGA A LOS PROFESORES: INCENTIVOS ECONÓMICOS EN ENTORNOS COMPLEJOS

Los estudios comparados suelen observar una asociación positiva entre los salarios que perciben los profesores en los distintos países (ajustados por PIB per cápita) y los resultados

de rendimiento que obtienen sus alumnos y colegios (Woessmann, 2005a; OECD, 2010; Dolton & Marcenaro-Gutierrez, 2011). No insistiremos, sin embargo, en los límites que presentan estos estudios a la hora de identificar relaciones de causalidad entre política y outcomes de interés.

Al mismo tiempo, el intento de identificar los efectos de las diferencias salariales entre docentes aprovechando posibles variaciones en esta variable dentro de un mismo país no deja de plantear problemas metodológicos importantes. Resulta difícil controlar la circunstancia de que las diferencias en el salario de los profesores pueden estar asociadas con su experiencia o cualificaciones, o incluso con las condiciones en que ejerce la docencia (por ejemplo, en algunos países es habitual que los profesores ocupados en colegios especialmente desfavorecidos reciban una remuneración superior a la del resto). No es fácil separar los posibles efectos de estas variables de los que puedan ser netamente atribuibles al nivel salarial de los docentes (Loeb & Page, 2000; Hanushek & Rivkin, 2007). Más aún, en esquemas variables de pago por resultados (performance-based pay), en los que el salario del profesor viene determinado por las mejoras en el rendimiento de sus alumnos, podríamos encontrarnos con un problema de causalidad inversa: no es que un mejor salario induzca a mejores resultados académicos, sino que es la consecución de mejoras de rendimiento lo que provoca mejoras salariales.

La superación de estas limitaciones vuelve a requerir la identificación de una fuente de variación exógena a las relaciones que se quieren estudiar. Y volveríamos a situar como diseños de evaluación más robustos los que cuentan con un mecanismo de asignación aleatoria al tratamiento; en el caso que nos ocupa, verse afectado por algún tipo de cambio en el salario. En este sentido, podríamos fijarnos en los experimentos dirigidos a evaluar la efectividad de diversos esquemas de incentivos económicos para el profesorado.

Una parte importante de estos experimentos se ha dirigido a probar distintas modalidades de pago por resultados, principalmente incentivos que denominaríamos de «expectativa de ganancia», según los cuales el profesor recibe un incremento salarial si consigue mejorar el rendimiento de sus alumnos en una cierta medida[13]. Según apunta la evidencia acumulada, mientras que estos esquemas pueden tener efectos positivos sobre el rendimiento de los alumnos en países en desarrollo económico (Glewwe, Ilias, & Kremer, 2010; Sundararaman & Muralidharan, 2011; Esther Duflo, Hanna, & Rya, 2012), en contextos desarrollados, sus impactos suelen ser nulos (Fryer, 2011b; Springer et al., 2011).

Otros experimentos evalúan programas de incentivos económicos para el profesorado no condicionados a resultados. Un ejemplo interesante es el que encontramos en el experimento realizado entre 2009 y 2011 por Mathematica Policy Research dirigido a evaluar la efectividad de un programa de incentivos económicos a la movilidad del profesorado más efectivo, programa conocido como Teacher Transfer Initiative – TTI (Glazerman, Protik, Teh, Bruch, & Max, 2013). Esta iniciativa, financiada por el Departamento de Educación de los Estados Unidos, consistía en incentivar el traslado de profesores con un destacado «valor añadido»[14] aa centros

de bajo rendimiento del mismo distrito, ofreciéndoles una bonificación de 20.000 dólares a cambio de trabajar allí durante al menos dos años. El experimento se llevó a cabo implicando a 115 colegios de primaria y secundaria situados en diez distritos escolares. El objetivo de la evaluación era estimar los impactos de la intervención sobre las dinámicas organizativas de los colegios y sobre el rendimiento académico de sus alumnos (en lectura y matemáticas), valorando al mismo tiempo la medida en que el programa consigue retener al profesorado trasladado al término de los dos años de duración del período de incentivos.

El diseño del experimento implicó, en primer lugar, la identificación de los centros de bajo rendimiento de los distritos seleccionados con plazas vacantes de profesorado por cubrir[15]. A continuación, los centros fueron emparejados de acuerdo con el curso y asignatura donde se localizaba la vacante y de acuerdo también con otras variables (puntuaciones académicas medias y proporción de alumnos desaventajados). Constituidos los bloques de parejas, un mecanismo de aleatorización se encargaba de distribuir a qué equipos docentes de cada bloque se abría la posibilidad de cubrir la vacante con un profesor TTI y a cuáles no (y, por tanto, tenían que cubrirlo a través de los canales convencionales de selección y contratación); los primeros son los equipos docentes «tratados» (85 equipos) y los segundos los equipos docentes «controles» (80 equipos).

En lo que respecta a los impactos organizativos, el experimento demuestra que el hecho de incorporar a un profesor TTI provoca cambios en la configuración de los equipos docentes de los cursos implicados; en concreto, los colegios tienden a equilibrar la composición de los equipos reasignando profesores menos experimentados a aquellos equipos donde se incorpora un profesor TTI. En cuanto a los resultados académicos, se constata el impacto positivo de la iniciativa sobre el rendimiento en lectura y matemáticas de los alumnos de primaria, sobre todo el segundo año del programa; durante el primer año, los impactos son únicamente positivos en el caso de los alumnos de los profesores de TTI y no para el conjunto de los alumnos de los equipos donde estos profesores se incorporan. En cambio, el programa no se muestra efectivo en los centros de secundaria. En lo que concierne a la ratio de retención del nuevo profesorado contratado, vía TTI o vía convencional, se constata que esta no difiere entre unos y otros al cabo de los dos años de duración del programa; en cambio, se muestra significativamente superior entre los profesores TTI al término del primer año de implicación en la iniciativa.

EN CATALUÑA...

- La definición de las condiciones salariales y de incentivación económica de los docentes es un elemento central de las políticas de profesorado, al lado de otras cuestiones como los sistemas de selección y acceso a la función docente, las condiciones de trabajo y carreras profesionales, la formación inicial y continua, los mecanismos de evaluación docente, el apoyo a la profesionalización y el liderazgo de los centros, etc. En Cataluña, buena parte de estos ámbitos han sido objeto de intervenciones y reformas a lo largo de los últimos años (Bonal & Verger, 2013).

- En el caso de los complementos e incentivos salariales del profesorado, podríamos referirnos a las medidas tomadas recientemente con relación a los denominados estadios de promoción. Así, el Acuerdo GOV/29/2012, de 27 de marzo, y la posterior Orden ENS/330/2014, de 6 de noviembre, establecen el procedimiento de promoción docente por estadios para el período 2012-2015, modificando la normativa de reconocimiento de sexenios vigente desde 1994. Según las disposiciones actuales, para superar un estadio se requieren diez créditos. De forma general, seis de estos créditos se obtienen por años de servicio activo (nueve créditos en el caso del primer estadio) y cuatro (uno en el primer estadio) por acumulación de determinados méritos. Uno de estos méritos remite a la implicación del docente en la mejora de los resultados del centro. En este punto, una de las medidas a considerar pasa por la asignación al profesorado de los resultados de las evaluaciones de los colegios (calificaciones ordinarias, pruebas de competencias, niveles de graduación y continuidad educativa, etc.).
- En concreto, se otorgará a cada profesor hasta un crédito por cada curso en que el centro obtenga una mejora significativa en su rendimiento o mantenga unos buenos resultados en sus indicadores. La superación de un estadio representa un incremento salarial mensual de entre 100 y 130 euros, un importe que se añade a otros complementos retributivos (por destinación y específicos, y trienios docentes).
- Este sistema de méritos introduce una cierta lógica de pago por resultados en las condiciones retributivas del profesorado. No obstante, no podemos decir que su diseño se haya fundamentado en ningún ejercicio riguroso de evaluación *ex-ante* (por ejemplo, revisión de evidencias o fundamentación de medidas de efectividad docente), al tiempo que su formulación deja poco margen al planteamiento de evaluaciones retrospectivas de impacto.

3.4 LAS TIC COMO HERRAMIENTA DE ENSEÑANZA-APRENDIZAJE: ACCESS TO ALGEBRA I

La literatura internacional sobre el uso de las TIC como herramienta de aprendizaje suele atribuirle impactos generales nulos o modestos. En cambio, la aplicación de las TIC incrementa notablemente su capacidad de impacto cuando forma parte de una estrategia pedagógica globalmente planificada y cuando es supervisada por tutores o docentes (Cheung & Slavin, 2012a, 2012b; Hew & Cheung, 2013).

Fijémonos, por ejemplo, en la evaluación que Heppen et al. (2012) hace del programa Access to Algebra I. Este programa consiste en el desarrollo de un curso de álgebra en línea basado en tres recursos básicos: contenidos prácticos interactivos y software de seguimiento y autoevaluación; profesorado especialista a distancia y responsables de centros encargados de la supervisión de los alumnos y del contacto con profesores y familias. La prueba piloto de este programa se planteó siguiendo un diseño experimental. Se identificaron 68 centros candidatos a desarrollar el programa, centros de secundaria inferior (middle schools) ubicados en áreas rurales de los estados de Maine y Vermont. Ninguno de estos centros ofrecía una asignatura específica de álgebra en el último curso de esta etapa educativa (8.º curso), es decir, los contenidos básicos de álgebra se impartían en el marco de la asignatura ordinaria de matemáticas. La mitad de estos centros fueron seleccionados aleatoriamente para desarrollar el programa Access to Algebra I como parte del currículo de sus estudiantes de 8.º curso,

mientras que la otra mitad (centros control) siguieron desarrollando su curso presencial tradicional de matemáticas. El estudio compara los progresos académicos de aquellos alumnos que, antes del momento de la aleatorización, habían sido identificados por el profesorado de unos centros y otros como alumnos aptos para recibir el programa, y se fija en dos medidas de outcome: puntuaciones de los alumnos en pruebas estandarizadas de álgebra y otras competencias matemáticas realizadas a final de curso, y nivel de seguimiento de la asignatura de álgebra en cursos posteriores (hasta 10.º curso). Pues bien, la evaluación demuestra que los alumnos que participan en el programa obtienen un nivel competencial en álgebra sensiblemente superior a los alumnos elegibles de los centros control, que esta ganancia no se produce a expensas de otros aprendizajes matemáticos no algebraicos (es decir, los alumnos del programa alcanzan el mismo nivel de matemáticas convencionales que los alumnos que las estudian en los centros control), y que todo ello estimula a los estudiantes participantes para seguir cursando la asignatura de álgebra en cursos más avanzados.

EN CATALUÑA...

- En el curso 2011-2012, el Departament d'Ensenyament de la Generalitat congeló el programa EduCat 1x1, iniciado dos cursos antes, de digitalización de las aulas catalanas de secundaria y del ciclo superior de primaria. El programa contemplaba la provisión al alumnado de ordenadores portátiles como herramienta personal de aprendizaje y trabajo. En el momento en que fue paralizado, el programa estaba activo en más de seiscientos centros. Ni la apuesta por el programa ni la decisión de congelarlo se apoyaban en evidencia empírica alguna sobre su capacidad de impacto en el proceso de aprendizaje de los alumnos.

- Por otra parte, desde el curso 2013-2014, el Departament d'Ensenyament participa en la implantación del programa mSchools, desarrollado desde el centro mEducation del Mobile World Capital Barcelona. El programa mSchools cuenta con la financiación de la fundación GSMA, por valor de 2,5 millones de euros hasta 2018. El programa quiere propiciar un despliegue curricular completo vinculado a la tecnología móvil y a sus aplicaciones en el entorno social y económico, y se plantea como objetivo de fondo contribuir a mejorar los resultados académicos de los alumnos. Una de las medidas estrella de este programa consiste en la introducción de la asignatura optativa Mobilitzem la informàtica (Movilicemos la informática) en el currículo de 4.º de la ESO. En esta asignatura los alumnos aprenden, de forma cooperativa y con la ayuda de mentores en línea, a diseñar aplicaciones móviles. En el curso 2014-2015, más de 11.000 alumnos cursarán la asignatura. Actualmente desconocemos la efectividad de Mobilitzem la informàtica o del programa mSchools en su conjunto a la hora de lograr los resultados educativos que se plantean. Conocer sus impactos debería permitir dimensionar mejor su escalabilidad apostando por aquellos componentes de innovación que pudieran mostrarse más efectivos y refinando aquellos que lo fueran menos.

3.5 LAS TUTORÍAS INDIVIDUALIZADAS: TIME TO READ Y SWITCH-ON READING

Hablamos ahora de un mecanismo específico de atención a la diversidad: las tutorías individualizadas o 1x1. Este instrumento persigue mejorar el rendimiento competencial de los alumnos más desaventajados a través de un refuerzo intensivo y personalizado al margen del grupo-clase ordinario. Los programas de tutorización individual pueden ser diversos,

dependiendo de si se realizan en horario lectivo o extraescolar, del perfil del personal tutor (profesores especialistas o no especialistas, paraprofesionales, voluntarios, etc.), de las materias trabajadas, del perfil y edad del alumnado atendido, de la frecuencia y duración de exposición al programa por parte de los alumnos (dosificación), etcétera.

Algunos de estos programas han sido evaluados experimentalmente. Por ejemplo, en el ámbito del refuerzo en lectura, podríamos hacer referencia a la prueba piloto experimental de los programas Reading Recovery (desarrollados por profesores especialistas) (May et al., 2014), Experience Corps (con tutores voluntarios mayores de 55 años) (Lee, Morrow-Howell, Jonson-Reid, & McCrary, 2012) o Reading Partners (con tutores voluntarios de perfil diverso) (Jacob, Smith, Willard, & Rifkin, 2014), todos ellos desarrollados en los Estados Unidos y dirigidos al alumnado de primaria con déficit en esta competencia. También en el Reino Unido encontramos ejemplos de evaluaciones experimentales de este mismo tipo de intervenciones. Es el caso de los programas Time to Read y Switch-on Reading.

En el año 2014, Time to Read estaba implantado en 100 colegios de primaria de Irlanda del Norte e intervenía sobre más de 1.200 alumnos de entre 8 y 9 años con déficit en competencia lectora. Mentores voluntarios se emparejan con estos alumnos y trabajan siguiendo un esquema de tutorización 1x1 que se desarrolla a lo largo de un curso escolar. Hasta el año 2008, las sesiones de tutoría, de 30 minutos de duración, tenían lugar una vez a la semana y al margen del grupo-clase de referencia. Un primer piloto experimental del programa, desarrollado entre septiembre de 2006 y junio de 2008 por investigadores del Centre for Effective Education (Queen's University Belfast), arrojó unos resultados decepcionantes en la mayoría de los outcomes considerados (básicamente variables relacionadas con la competencia lectora, las actitudes hacia la escolarización y la autoestima) (Miller & Connolly, 2012). A partir de 2008, y por indicación del propio informe de evaluación, el programa pasó a doblar la dosificación de las tutorías, incluyendo ahora dos sesiones semanales de 30 minutos cada una.

Entre octubre de 2008 y junio de 2010 tuvo lugar la segunda evaluación experimental del programa, llevada a cabo por el mismo equipo de evaluadores que se encargó de la primera (Miller, Connolly, & Maguire, 2012). Esta segunda prueba experimental contó con una muestra de 512 alumnos en 50 colegios de primaria. El procedimiento consistía en seleccionar aleatoriamente, dentro de cada colegio y entre los alumnos elegibles identificados por el profesorado, aquellos alumnos que recibirían el programa (un total de 263 alumnos) y aquellos que no lo recibirían (249 alumnos), con el objetivo de comparar los resultados correspondientes al cabo de dos años de seguimiento. Los resultados de esta segunda evaluación permitieron atribuir al programa impactos positivos en outcomes como la capacidad de descodificación, la velocidad y la fluidez en la lectura. En cambio, el programa no se mostraba efectivo en el ámbito de la comprensión lectora o a la hora de fomentar el gusto por la lectura o de incrementar la confianza en la eficacia lectora.

Por su parte, el programa Switch-on Reading ofrece a los alumnos participantes (jóvenes al inicio de la educación secundaria con un bajo nivel competencial en lectura) un mínimo de 40 sesiones de refuerzo individualizado de 20 minutos de duración cada una, realizadas fuera del grupo-clase ordinario y programadas por un período de diez semanas. La prueba piloto experimental de este programa contó con la financiación de la Education Endowment Foundation[19], y se llevó a cabo entre diciembre de 2012 y abril de 2013 siguiendo un diseño ligeramente distinto al convencional (Gorard, See, & Siddiqui, 2014). En este caso, se aleatoriza el momento de entrada al programa y no el hecho de recibirlo o no recibirlo. En otras palabras, nadie se queda sin participar en él, aunque unos entran antes que otros. Aquí, de los 300 alumnos elegibles para Switch-on Reading, se asignó aleatoriamente los 150 que entrarían en una primera fase (trimestre 1) y los 150 que se quedarían a la espera de entrar en una segunda fase (trimestre 2). A todos ellos se les administró una prueba estandarizada de comprensión lectora antes de la aleatorización y justo después de finalizar la primera fase del programa. Así pues, la estimación del impacto tiene en cuenta el primer trimestre de implementación del programa, período en el que los alumnos en espera actúan como control de los alumnos participantes en primera instancia.

Este diseño neutraliza posibles reservas éticas frente a la asignación aleatoria del recurso, y al mismo tiempo elimina eventuales efectos de desánimo entre los individuos control. La principal limitación del diseño es que no permite observar impactos más allá del corto plazo (primer trimestre de implementación del programa). Y hay que decir que los impactos detectados son muy positivos. En efecto, la comparación de las puntuaciones pre y post-programa de los individuos control y los individuos tratamiento indica que este es efectivo para el conjunto de alumnos participantes, particularmente para los alumnos académica y socialmente más vulnerables.

Las conclusiones de los experimentos Time to Read y Switch-on-Reading corroboran la capacidad de impacto que la literatura internacional atribuye a las fórmulas de tutorización 1x1[20]. Para que estos impactos puedan perdurar en el tiempo, se requiere que las tutorías partan de un diseño de actividades bien planificadas y coordinadas con los procesos de aprendizaje que tienen lugar en el grupo-clase ordinario (Borman et al., 2007; Correnti, 2009). Asimismo, la efectividad de la intervención se incrementa a medida que aumenta el tiempo y la frecuencia de realización de las tutorías, y cuando estas corren a cargo de profesorado cualificado o de personal bien formado y asesorado (Slavin, Lake, Davis, & Madden, 2011; Ehri, Dreyer, Flugman, & Gross, 2007).

EN CATALUÑA...

- Entendidas como instrumento compensatorio o de atención a la diversidad, las tutorías individualizadas o tutorías 1x1 están poco implantadas en la vida ordinaria de los centros en Cataluña. Hay que decir, sin embargo, que no son pocas las voces que, desde los colegios e institutos, señalan la conveniencia de llegar a articular este servicio, cosa que suele resultar imposible por falta de recursos humanos.
- Sí se han desarrollado, en cambio, programas de tutorización individual o en grupos reducidos, implementados dentro de los centros, pero en horario extraescolar, y en colaboración con entidades locales, fundaciones o universidades. Este es, por ejemplo, el caso del programa LECXIT: Lectura per a l'èxit educatiu (LECXIT: Lectura para el éxito educativo) iniciado en el curso 2011-2012, promovido por la Fundación "la Caixa", la Fundación Jaume Bofill y el Departament d'Ensenyament de la Generalitat de Cataluña, y dirigido a niños de 4.º a 6.º de primaria [16]; del programa Èxit 1: Reforç escolar i activitats complementàries (Éxito 1: Refuerzo escolar y actividades complementarias), iniciado el curso 2001-2002, impulsado por el Consorcio de Educación de Barcelona y dirigido al alumnado de 5.º y 6.º de primaria y 1.º y 2.º de la ESO[17]; o de los Talleres de Estudio Asistido, vinculados a los Planes Educativos de Entorno iniciados por el Departament d'Ensenyament en el curso 2004-2005 y dirigidos principalmente al alumnado de secundaria[18].
- Sin embargo, no sabemos mucho sobre el impacto neto de estos programas, lo cual nos permitiría valorar la conveniencia de potenciarlos y/o probar su traslado al marco ordinario de funcionamiento de los centros.

3.6 LA IMPLICACIÓN DE LAS FAMILIAS: MALLETE DES PARENTS Y READY4K!

Nos fijamos ahora en la evaluación de aquellos programas que tratan de incentivar la implicación de las familias en los procesos y aprendizajes escolares de los hijos. Son programas diversos y abarcan tanto actuaciones que persiguen involucrar a las familias en las actitudes de aprendizaje y en las actividades escolares que tienen lugar en el entorno doméstico como intervenciones que tratan de promover la participación de padres y madres en la vida cotidiana del colegio. Según la evidencia internacional, cuando lo que está en juego es la mejora del rendimiento y las actitudes de los alumnos, el primer tipo de actuaciones tiende a ser más efectivo que las destinadas a la participación escolar de las familias (Jeynes, 2005, 2007; van Steensel, McElvany, Kurvers, & Herppich, 2011).

Los principales metanálisis sobre la materia remarcan, sin embargo, que una gran mayoría de los estudios que se han ocupado de esta cuestión se basan en diseños de evaluación no del todo rigurosos (Jeynes, 2005, 2007; Mattingly, Prislin, McKenzie, Rodriguez, & Kayzar, 2002; Van Voorhis, Maier, Epstein, Lloyd, & Leung, 2013). Y, sin embargo, algunos de estos programas sí han podido ser evaluados mediante diseños experimentales.

Fijémonos en la evaluación de un programa de orientación llevado a cabo en colegios socialmente desfavorecidos de los alrededores de París. La actuación se enmarca en un

programa más amplio dirigido a potenciar la relación familia-colegio (Mallete des parents), si bien se implementa y evalúa de forma independiente. Consiste en la programación de dos charlas con padres y madres de alumnos en el último curso de la escolarización obligatoria, alumnos que el profesorado ha identificado como en riesgo de abandono. Las charlas están a cargo del director del centro y tienen un componente tanto informativo como de orientación: se remarca la importancia de la educación y de escoger bien el itinerario de continuidad. En cada charla participa una media de diez padres o madres. La actuación cubre una muestra de 37 colegios voluntarios. A través de un sorteo se escoge qué clases dentro de estos colegios participarán en el programa (recibirán las charlas). Los alumnos de las clases no seleccionadas (en colegios no seleccionados) conforman el grupo de control. A partir de aquí, se comparan los niveles de repetición, abandono y rendimiento de los alumnos tratados y de los alumnos control a lo largo de dos cursos escolares y se mide si las medidas de unos y otros son significativamente diferentes. Los resultados obtenidos acaban identificando impactos significativos y duraderos del programa sobre los outcomes educativos clave: los hijos de las familias participantes se mantienen más tiempo que los alumnos control en el sistema educativo (principalmente en opciones menos demandadas de formación profesional), y durante este tiempo rinden más y repiten menos (Goux, Gurgand, & Maurin, 2013).

Igualmente interesante en este sentido es la prueba piloto experimental de READY4K!, en la ciudad de San Francisco, un programa que pretende inducir mejoras en la competencia lectora de los alumnos de educación infantil y que se basa en el envío de mensajes telefónicos SMS a sus familias. El experimento en cuestión tuvo lugar durante el curso 2013-2014 en 31 centros preescolares del principal distrito escolar de la ciudad. Entre las 519 familias con hijos escolarizados en estos centros que, al empezar el curso, mostraron interés en participar en el piloto, se designó aleatoriamente a aquellos que recibirían el protocolo de SMS del programa y a los que recibirían SMS de tipo «placebo». El protocolo de SMS de READY4K! incluye el envío de tres mensajes de texto semanales durante ocho meses del calendario lectivo. Los contenidos de los mensajes combinan consejos para la práctica de la lectura con los hijos, datos sobre la importancia de los aprendizajes en la primaria infancia y textos de ánimo a la tarea educativa de la familia en el ámbito doméstico. Por el contrario, los SMS «placebo» se envían a las familias que actúan como grupo de control a razón de uno cada dos semanas, e incluyen contenidos generales sobre los procedimientos de inscripción en la educación primaria o sobre el calendario de vacunaciones, entre otros. Los resultados de la evaluación muestran que el protocolo de SMS de READY4K! incrementa la implicación de padres y madres en la práctica lectora de los hijos en el ámbito doméstico, así como su participación en la vida escolar. A su vez, se constata que estos incrementos revierten en mejoras significativas en determinadas áreas de la competencia lectora de los niños (por ejemplo, conocimiento del alfabeto y fonética). Más aún, los beneficios del programa son particularmente relevantes entre las familias negras e hispanas y los hijos de estas.

EN CATALUÑA...

- Cataluña dispone de una larga experiencia en el desarrollo de actuaciones dirigidas a fomentar la implicación de las familias en los procesos escolares de sus hijos, actuaciones que pueden ser, y han sido, muy diversas: acciones formativas (para que las familias conozcan el funcionamiento de los centros, la estructuración del sistema educativo, determinados recursos de educación no formal, etc.), de sensibilización[21] (dirigidas a concienciar a las familias sobre la importancia de la educación y sobre la necesidad de apoyar el progreso escolar de los hijos), de capacitación (intentando que padres y madres dispongan de las competencias necesarias para acompañar el aprendizaje de sus hijos), de participación (aproximando a padres y madres a los canales de participación familiar en la vida escolar), etc. Y son también diversos los agentes responsables de estas iniciativas: los centros escolares en sí, a través de sus tutores y equipos directivos; los ayuntamientos, a través de sus programas y acciones familia-colegio; fundaciones y entidades del tercer sector, como proveedores de intervenciones públicas en el marco de iniciativas filantrópicas, etc.

- En cualquier caso, no son pocas las voces que, desde la comunidad educativa y desde el mundo académico, insisten en la conveniencia de continuar apostando por medidas que involucren a las familias en la escolarización de los hijos, tanto en el ámbito doméstico como en el de la participación escolar (Collet & Tort, 2011; Comas, Escapa, & Abellán, 2014). La defensa de esta apuesta, no obstante, ganaría fuerza si pudiéramos demostrar, de forma robusta, la efectividad de los programas familia-colegio en Cataluña o si estuviéramos al menos en disposición de incorporar los aprendizajes de eventuales evaluaciones de impacto —inexistentes hoy en día— en su diseño o reforma.

Notas:

[9] *PISA = Programme for International Student Assessment (OCDE); TIMSS = Trends in International Mathematics and Science Study (IEA); PIRLS = Progress in International Reading Literacy Study (IEA).*

[10] *Prácticas selectivas dirigidas bien a captar al alumnado de perfil académico o socioeconómico más elevado, bien a evitar la entrada del alumnado académica o socioeconómicamente más vulnerable.*

[11] *Los autores de estas evaluaciones atribuyen el éxito de las charter de Boston y Nueva York al modelo educativo No Excuses que desarrolla buena parte de ellos. Este modelo apuesta por un alto nivel de exigencia y de expectativas en el ámbito de las actitudes y de los resultados académicos de los alumnos, al tiempo que opta por la extensión del horario lectivo y por sistemas de tutorización y monitorización más estrechos e informados (Carter, 2000; Thernstrom & Thernstrom, 2004; Whitman, 2008).*

[12] *Una posible explicación del corto alcance temporal de los impactos cognitivos de programas como los que hemos mencionado aquí se encuentra en la baja calidad de los centros educativos a los que los niños acceden después de participar en ellos, y contra la que los programas en cuestión no consiguen «inmunizar» (Currie & Thomas, 2000; Loeb & Lee, 1995). Por su parte, Heckman et al. (2012) sostienen que los impactos de estos programas que son observables en la edad adulta se deben principalmente a los cambios que estos consiguen en el terreno actitudinal y motivacional de los niños.*

[13] *El experimento realizado por Fryer et al. (2012) con 150 profesores de nueve colegios de Chicago Heights permite comparar el impacto de esquemas de «expectativa de ganancia» con la efectividad de un esquema de incentivos basada en el «miedo a la pérdida» (loss aversion), según el cual los profesores son bonificados de antemano, y obligados posteriormente a devolver parte de la bonificación si al final del período estipulado no se consiguen los resultados esperados. El estudio concluye que el esquema de miedo a la pérdida, sobre todo cuando se define de acuerdo con resultados-ganancias colectivos, funciona mejor que el esquema tradicional de incentivos individuales.*

[14] *Las medidas de «valor añadido» del profesorado tratan de describir cómo contribuyen los profesores a los cambios en el rendimiento académico de los alumnos, manteniendo constantes otras variables que escapan a su control, por ejemplo, la extracción socioeconómica de los alumnos o su rendimiento previo (McCaffrey et al. 2004; Lipscomb et al. 2010). En cuanto al experimento que nos ocupa, se consideran elegibles para la TTI aquellos profesores cuya puntuación de «valor añadido» les situaba entre el 20 % de profesores más efectivos del distrito.*

[15] *Se consideran de bajo rendimiento los colegios que se encuentran entre el 20 % de centros con peor nivel académico del distrito.*

[16] *Más información en http://www.lectura.cat/*

[17] *Más información en http://www.edubcn.cat/exit/reforc_escolar_i_activitats_complementaries*

18 *Más información en http://ves.cat/hrLH*

19 *Financiación recibida en el marco de la primera convocatoria competitiva de ayudas a proyectos de esta institución (año 2012).*

20 *Las evaluaciones experimentales de otros programas de tutorización individualizada apuntan en esta misma dirección. Véanse, por ejemplo, las conclusiones de la evaluación del programa 1x1 Catch Up Numeracy, para alumnos de educación primaria con déficits en matemáticas (Rutt, Easton, & Stacey, 2014), o los resultados del programa Becoming a Man, centrado en el trabajo conductual con niños de secundaria en riesgo de abandono escolar (Cook et al., 2014).*

21 *Ver, por ejemplo, el programa Familia y escuela en http://familiaiescola.gencat.cat/ca/*

4. LOS RETOS DE LA EVALUACIÓN CUASIEXPERIMENTAL: EJEMPLOS

Como ya hemos indicado más arriba, son diversos los métodos cuasiexperimentales que pueden utilizarse para evaluar los impactos de programas educativos. Si bien estos métodos no neutralizan el sesgo de selección de forma tan clara como el diseño experimental, bajo determinadas circunstancias algunos de ellos pueden ofrecer estimaciones de impacto ciertamente robustas. Se trata, principalmente, de la regresión discontinua, las variables instrumentales y los modelos de dobles diferencias. A continuación ilustramos su potencialidad en ámbitos de intervención relevantes para la agenda educativa catalana.

4.1 DISEÑOS DE REGRESIÓN DISCONTINUA

4.1.1 DE QUÉ SIRVE REPETIR CURSO

No es fácil identificar los efectos atribuibles al hecho de repetir un determinado curso. Al igual que suele ocurrir con buena parte de la política de evaluación de alumnos, nos encontramos ante una decisión evaluativa (no promocionar a un alumno) muy idiosincrática. Dentro del mismo sistema educativo, aun existiendo criterios comunes en cuanto a la aplicación de este instrumento, podemos acabar encontrando interpretaciones y usos de la repetición muy diversos entre colegios. Por ejemplo, puede ocurrir perfectamente que determinados centros tiendan más que otros a evitar la práctica de la repetición por motivos que pueden ser variados: pedagógicos (confianza en que la repetición no ayuda al progreso escolar de los alumnos), económicos (ahorro de los costes asociados a la repetición) o de imagen (de efectividad del centro y/o de buen perfil del alumnado).

En resumen, la evaluación de los impactos de la repetición de curso no puede basarse en la comparación de los resultados a corto, medio o largo plazo entre los alumnos que repiten un determinado curso y los que lo superan. Necesitamos contar con métodos que, de entrada, eliminen los sesgos entre las características de repetidores y no repetidores. Siendo difícil aplicar un diseño experimental a la evaluación de la repetición, algunos estudios se han aproximado mediante distintos diseños de regresión discontinua.

El estudio de Jacob y Legren (2009) ofrece un ejemplo interesante en este sentido. Los autores aprovechan la introducción en el año 1997, en los colegios públicos de Chicago, de un test de competencias externo y estandarizado dirigido a determinar la promoción o repetición de curso de los alumnos en momentos clave de su itinerario escolar. En particular, mediante una aproximación de regresión discontinua, los autores comparan a) los niveles de graduación al finalizar la educación secundaria de aquellos alumnos que repitieron curso en 6.º o en 8.º (colegio elemental) obteniendo en el test de competencias una puntuación justo por debajo de la mínima exigida para pasar de curso y b) los niveles de los alumnos de las mismas cohortes que promocionaron obteniendo una puntuación justo por encima del umbral mínimo requerido[22]. De

esta manera demuestran que el hecho de repetir 6.º no afecta a la probabilidad de finalizar con éxito la secundaria; en cambio, repetir 8.º sí tiene un impacto negativo sobre esta probabilidad, particularmente entre las chicas afroamericanas.

Dos estudios de Marco Manacorda utilizan este mismo método para evaluar los efectos de la repetición de curso en Uruguay (2008, 2012). En uno de estos estudios (Manacorda, 2012), el autor aprovecha el hecho de que en este país se instauró una normativa en el año 1996 que establecía la repetición de curso automática de los alumnos de secundaria con más de tres asignaturas suspendidas en la evaluación final[23]. Aprovechando esta circunstancia, y recurriendo a un modelo de regresión discontinua[24], se comparan los resultados educativos de los alumnos que se encuentran a un lado y otro del umbral de las tres materias pendientes, considerando que el hecho de situarse justo por encima de este punto de corte (y repetir) o justo por debajo (y promocionar) no refleja diferencias relevantes en las características iniciales de unos alumnos y otros. De esta manera, el autor muestra cómo el hecho de repetir reduce significativamente el tiempo que los alumnos permanecen en el colegio y, por tanto, sus niveles de graduación.

Los resultados de las evaluaciones a las que nos hemos referido se encuentran en sintonía con las conclusiones a las que llegan las principales revisiones sistemáticas de la literatura sobre esta cuestión (Allen, Chen, Willson, & Hughes, 2009; Holmes, 1989; Jimerson, 2001), a saber: que la decisión de hacer repetir a un alumno no es solo una decisión muy costosa desde el punto de vista económico, sino que además no tiene impactos significativos (o de detectarse alguno, es negativo) ni sobre las actitudes ni sobre el progreso educativo de los alumnos repetidores.

EN CATALUÑA...

- La política de repeticiones es, en Cataluña y España, una práctica bastante más extendida que en muchos países de nuestro entorno, tanto en la educación primaria como en la secundaria (Dupriez, Dumay, & Vause, 2008; Goos et al., 2013).
- Sin ir más lejos, desde la entrada en vigor de la Ley Orgánica General del Sistema Educativo (LOGSE, de 4 de octubre de 1990), se han aplicado diversas medidas para regular el uso de la repetición por parte de los centros. Se ha apostado por sistemas de promoción automática entre cursos (con posibilidad de repetir al finalizar ciclos) y también por la repetición de cualquier curso a partir de un determinado número de asignaturas suspendidas. Paralelamente, han sido también diversas y cambiantes las vías abiertas para la recuperación de asignaturas pendientes (recuperaciones durante el curso o en septiembre).
- En Cataluña, el modelo actual en las enseñanzas secundarias obligatorias establece la repetición de curso para los alumnos con tres o más materias suspendidas en la evaluación final extraordinaria propia de cada curso (Orden EDU/295/2008). De forma excepcional, si el equipo docente así lo considera o si el alumno ha agotado el número máximo de repeticiones permitidas en la etapa (dos), se podrá eximir de la repetición a alumnos con tres o más asignaturas suspendidas. Por su parte, a partir del curso 2011-2012, la Orden ENS/56/2012 modifica el calendario de la evaluación final extraordinaria, restaurando los exámenes de septiembre.

- En cualquier caso, conviene destacar que ninguno de estos modelos ha sido contrastado empíricamente, y menos aún evaluado en términos de sus posibles impactos sobre outcomes educativos clave, que son, principalmente, la mejora y recuperación del rendimiento y de los logros formativos de los alumnos repetidores.

4.1.2 BECAS Y AYUDAS A LOS ESTUDIANTES

Las ayudas directas que las administraciones, o algunas entidades filantrópicas, conceden a los estudiantes con el objetivo de facilitar su escolaridad son un instrumento clave de política educativa. No obstante, evaluar el impacto que puede tener este instrumento sobre las oportunidades educativas de los estudiantes no es siempre una tarea sencilla.

En el plano de la comparativa internacional, suele ponerse de manifiesto que no existe una relación clara entre el nivel global de gasto en educación (medida en referencia al PIB) y los resultados educativos medios de los países. Más aún, la debilidad de esta asociación es especialmente evidente entre los países desarrollados (OECD, 2014). Todo parece indicar que, a partir de un determinado umbral, la efectividad del gasto educativo se convierte más en una cuestión de estrategia (cómo se gasta) que de volumen (cuánto se gasta). En cambio, algunos estudios han puesto de manifiesto la existencia de una asociación positiva entre, por una parte, el peso que tiene el gasto en ayudas a los estudiantes sobre el total del gasto en enseñanzas no universitarias y, por otro, sus logros agregados (niveles formativos superados) (Alegre & Benito, 2014).

Una vez más debemos ser prudentes en la interpretación de estos resultados, no solamente por los problemas de endogeneidad que plantean, sino también por la gran diversidad de esquemas de ayudas que suelen reunirse dentro de una única variable: becas y premios de estudio, subsidios directos a las familias para costear el gasto privado derivado de la escolarización, préstamos preferenciales, etc. De igual forma, la comparativa entre países no siempre consigue distinguir entre las distintas modalidades de acceso a estas ayudas, en función de méritos escolares (modelo merit-based) y/o de determinadas medidas de carencia económica (modelo need-based).

Incluso cuando el interés se concentra en un determinado programa de ayudas (desarrollado en un único territorio), se plantea el reto de identificar grupos de comparación creíbles, es decir, estudiantes no beneficiarios verdaderamente comparables a los estudiantes que efectivamente reciben la ayuda en cuestión. Son numerosos los sesgos que pueden producirse entre beneficiarios y no beneficiarios: sesgos competenciales en los modelos de ayuda merit-based; o sesgos socioeconómicos y culturales en los modelos need-based. Incluso si pudiéramos controlar la incidencia de estos sesgos, podría ocurrir perfectamente que beneficiarios y no beneficiarios difirieran en variables no observables relacionadas con los outcomes de interés[25].

Una manera de superar estas limitaciones pasaría por sacar partido de los umbrales impuestos a los criterios de elegibilidad que definen buena parte de los programas de becas escolares, mediante la aplicación de diseños de regresión discontinua. En el caso de aquellas becas cuya concesión exige haber aprobado un mínimo de asignaturas o haber obtenido unas calificaciones o puntuaciones académicas mínimas, se trataría de comparar lo que les pasa a los estudiantes que son beneficiarios de ellas y que están justo por encima del mínimo exigido de asignaturas aprobadas o de calificaciones obtenidas (grupo de tratamiento) con lo que les sucede a los estudiantes no beneficiarios que se encuentran justo por debajo de los umbrales académicos en cuestión (grupo de comparación). En el caso de las becas que se otorgan por renta, se compararían los alumnos becados cuyos ingresos familiares se encuentran por debajo del máximo exigido (grupo de tratamiento) con los alumnos con rentas familiares justo por encima de este mismo máximo (grupo de comparación). En cualquier caso, es de esperar que unos alumnos y otros, tanto los que están justo por encima como los que están justo por debajo de estos umbrales académicos o de renta, se parezcan no solamente en aquellas características observables que podemos controlar, sino también en aquellas no observables que pueden estar relacionadas con la variable de asignación.

Este procedimiento se ha utilizado en la estimación de impactos de distintos programas de becas universitarias (basados en rendimiento o en renta), principalmente en los Estados Unidos (DesJardins, McCall, Ott, & Kim, 2010; Rubin, 2011; Scott-Clayton, 2011; Leeds & DesJardins, 2012; Curs & Harper, 2012; Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012). En este contexto, la evidencia acumulada tiende a concluir que, mientras que las ayudas basadas en la renta favorecen principalmente el acceso a la universidad, las becas basadas en el rendimiento incrementan las tasas de permanencia y graduación. En conjunto, la capacidad de impacto de estos dos tipos de becas se amplifica a medida que aumenta su importe y que se simplifican los procedimientos administrativos para su solicitud (Dynarski & Scott-Clayton, 2013).

El ejemplo que aquí exponemos remite a la evaluación realizada por Fack y Grenet (2013) sobre el programa francés Bourses sur Critères sociaux, más concretamente, sobre sus impactos en los niveles de matriculación, permanencia y graduación de los estudiantes que se benefician de él. Este es el principal programa de becas universitarias en Francia, basado en un esquema need-based. En el año 2010, una tercera parte de los universitarios franceses recibían algún tipo de ayuda procedente de este programa. En concreto, el programa define seis niveles diferentes de beca, a los que se accede en función de umbrales de «necesidad» (establecidos en base a la renta familiar, el número de hermanos y la distancia del hogar a la universidad). El nivel 1 comporta el pago de los gastos de matrícula y una transferencia de 1.500 euros anuales[26]. A partir de aquí, el incremento medio de los sucesivos niveles de la ayuda corresponde a un importe de 600 euros anuales. Aprovechando la exogeneidad de la definición de los distintos umbrales de necesidad y utilizando un diseño de regresión discontinua, los autores comparan los resultados de los estudiantes que solicitaron beca los años 2008 y 2009 y que acabaron encontrándose justo por encima y justo por debajo de los distintos umbrales que concedían acceso a los distintos niveles de ayuda.

Mediante este procedimiento, Fack y Grenet (2013) concluyen que recibir una ayuda de 1.500 euros al año (beca de nivel 1) incrementa hasta cuatro puntos porcentuales la probabilidad de acceder a la universidad y de permanecer en ella dos cursos después; entre los estudiantes de máster se observan impactos de una magnitud similar sobre la probabilidad de acabar graduándose. Sucesivos incrementos en la dotación de la ayuda (del nivel 2 de la beca en adelante) no reportan impactos positivos significativos sobre los niveles de matriculación universitaria, y tampoco se aprecian efectos diferenciales de las ayudas en función del sexo de los estudiantes o de su expediente académico preuniversitario.

EN CATALUÑA...

- Hablamos de programas de becas de índole muy diversa: desde las ayudas al alumnado con necesidades educativas especiales, pasando por las becas de comedor o para la adquisición de material escolar, hasta las becas generales para enseñanzas secundarias posobligatorias o para estudios superiores. En Cataluña, la gran mayoría de ayudas económicas que se conceden en las etapas educativas obligatorias tienen la renta familiar como principal criterio de prioridad. Este mismo esquema caracteriza buena parte de las becas existentes en la secundaria posobligatoria y en los estudios universitarios, aunque aquí convive con otras ayudas basadas en el rendimiento académico.
- El caso de las ayudas públicas en la universidad es especialmente paradigmático. Desde el curso 2011-2012, en Cataluña se programan las becas Equitat y las becas Excel·lència, ambas para estudiantes de grado y de enseñanzas no adaptadas al Espacio Europeo de Educación Superior. Las becas Equitat financian hasta un 50 % del precio público de la matrícula, según importes que se modulan por tramos de renta. Las becas Excel·lència, en cambio, financian estancias de movilidad internacional a estudiantes con un alto rendimiento académico.
- A día de hoy disponemos de poca evidencia sobre la capacidad que tienen los distintos programas de becas para mejorar las oportunidades de sus beneficiarios en las distintas etapas educativas. Uno de los pocos intentos de aproximación a esta cuestión lo tenemos en el estudio de Mediavilla (2012, 2013). Este autor evalúa los efectos de las becas y las ayudas al estudio implementadas en España en el período 2005-2006. Utilizando como referencia la Encuesta de Condiciones de Vida (2006), el autor concluye: los alumnos que han recibido una beca durante la enseñanza obligatoria tienen entre un 22 y un 25 % más de posibilidades de finalizar con éxito la educación posobligatoria que otros alumnos de características similares que no han sido beneficiarios de becas; esta probabilidad se incrementaría entre las chicas y los alumnos procedentes de hogares desfavorecidas[27].

4.2 EL USO DE VARIABLES INSTRUMENTALES

4.2.1 LA IMPORTANCIA DE LA RATIO DE ALUMNOS POR CLASE

El número de alumnos por clase que definen los centros y/o las normativas marco correspondientes es uno de los aspectos de la organización docente y escolar que ha acaparado tradicionalmente más atención académica y política. Ahora bien, determinar los efectos que este factor puede llegar a representar sobre el progreso académico de los alumnos tiene sus complicaciones.

En efecto, la simple comparación de los resultados que obtienen los alumnos en clases con una ratio más o menos elevada - ya se comparen alumnos dentro del mismo centro, alumnos entre centros o entre países- plantea serias limitaciones metodológicas. Podemos encontrar, por ejemplo, que los colegios intenten tratar en clases reducidas precisamente al alumnado con peor rendimiento, o que las autoridades educativas opten por ofrecer a los colegios desfavorecidos recursos docentes suplementarios que permitirían al conjunto del centro ampliar el número de clases reduciendo su ocupación. En este caso, la asistencia a grupos-clase reducidos podría asociarse con malos resultados, no tanto por el nivel de ocupación del aula, sino por el perfil de esta ocupación. Pero también podríamos encontrarnos con la situación contraria: las familias con más recursos y más preocupadas por la escolarización de sus hijos, o bien escogen estos centros que priorizan los grupos-clase reducidos, o bien presionan a los centros escogidos para que procedan de esta manera (Schlotter et al., 2010). En este caso, la eventual asociación entre clases reducidas y buenos resultados se explicaría por la extracción privilegiada de los alumnos que tenderían a concentrarse en ellas.

Distintos estudios han tratado de superar estas limitaciones mediante estrategias metodológicas diversas. Dos de estas estrategias, ambas basadas en el uso de variables instrumentales, merecen una breve explicación[28]. En primer lugar, se han aprovechado las fluctuaciones en el momento del nacimiento como fuente de variación exógena del volumen de alumnos que acaban componiendo los distintos cursos y, por tanto, las aulas que los conforman. De este modo, se entiende que el hecho de nacer antes o después del mes de corte que define la adscripción a un curso o a otro no depende de factores relacionados con el outcome de interés (Hoxby, 2000; Woessmann & West, 2006). Cuando en los centros se detectan diferencias relevantes entre el número de alumnos que componen un determinado curso y el número de alumnos que componen el curso adyacente, cuando se excluye la posibilidad de que estas diferencias sean debidas a dinámicas de movilidad de los estudiantes y cuando estas diferencias se traducen en un mismo número de clases por curso (unas más nutridas, otras menos), entonces las diferencias de rendimiento entre los grupos de ambos cursos pueden ser atribuidas a las diferencias de tamaño de las clases respectivas.

En segundo lugar, algunos estudios han utilizado el establecimiento de normativas sobre el número máximo de alumnos por clase como instrumentos del tamaño de las aulas (Angrist & Lavy, 1999; Woessmann, 2005a). La aplicación de este diseño se justifica porque no existe una relación lineal entre el número de alumnos matriculados en un curso y el número de alumnos de las clases que lo componen. A partir de aquí, se explotan las discontinuidades que, dentro de cada centro, se producen en torno al umbral máximo de alumnos por aula, y se compara lo que les ocurre a los alumnos en las aulas de máxima ocupación con lo que les ocurre a los alumnos que, para no sobrepasar el umbral máximo, son asignados a aulas subocupadas del mismo curso.

Una visión general sobre los resultados que obtienen ambas aproximaciones cuasiexperimentales nos llevaría a afirmar que los efectos del tamaño de las clases sobre los logros de los alumnos tienden a ser más bien modestos. No obstante, en algunos casos y bajo ciertas circunstancias, esta práctica organizativa sí consigue efectos notables. Algunas revisiones sistemáticas sobre la evidencia acumulada en torno a esta cuestión indican, por ejemplo, que las clases reducidas (orientativamente, con menos de quince alumnos) pueden ser particularmente efectivas para los alumnos más desfavorecidos, con peor rendimiento académico y de menos edad, y siempre y cuando se cuente con un profesorado preparado para sacarles provecho (McGiverin, Gilman, & Tillitski, 1989; Slavin, 1989; Hattie, 2005).

EN CATALUÑA...

- El Ministerio de Educación, Cultura y Deporte dictó para el curso 2012-2013 un incremento general de las ratios máximas de alumnos por aula en la etapas de primaria (pasando de una ratio máxima de 25 a 30 alumnos) y en secundaria (de 30 a 35 alumnos). Este incremento se justificaba como una medida de ahorro que no tenía por qué poner en riesgo la calidad del sistema y las oportunidades educativas de los alumnos. Desde el Departament d'Ensenyament de la Generalitat se denunció la posible invasión de competencias que esta orden representaba y se optó por mantener en la mayoría de colegios e institutos la misma ratio máxima que el curso anterior. En Cataluña, en el curso 2014-2015, la ratio de alumnos por aula continúa estableciéndose en 25 alumnos en los colegios públicos (educación infantil y primaria) y 30 en los institutos (ESO).
- En definitiva, motivados por unas razones u otras, los cambios en la regulación de estas ratios han sido frecuentes en Cataluña y en España en las últimas décadas. Sin embargo, en ningún caso estos cambios se han apoyado en evidencias empíricas sobre sus posibles repercusiones en el rendimiento de los alumnos; menos aún, por tanto, en estimaciones económicas coste-efectividad o coste-beneficio de su aplicación.

4.2.2 LA LIBERTAD DE ESCOGER CENTRO EDUCATIVO

Nos preguntamos aquí sobre los impactos que las políticas de apertura del margen de elección escolar pueden producir en términos de logros y oportunidades educativas. En opinión de algunos autores, estas políticas representan uno de los principales motores de buena parte de las reformas educativas planteadas a escala mundial (Burch, 2009; Ball, 2012).

Estudios basados en datos de los programas PISA, TIMMS o PIRLS, suelen observar una asociación positiva entre, por un lado, el margen de elección escolar de los países y, por otro, el nivel medio y también de equidad de los resultados que obtienen sus estudiantes (Schütz et al., 2007; Woessmann, Luedemann, Schuetz, & West, 2007). En el plano individual, sin embargo, cuando estas asociaciones son corregidas por la extracción socioeconómica y cultural de los alumnos y de los colegios donde acceden, el hecho de estudiar o no en un centro de libre elección suele perder toda su relevancia (OECD, 2014).

No obstante, estos estudios presentan limitaciones a la hora de controlar la endogeneidad existente entre la elección escolar y los resultados educativos[29]. La única manera de superar estas limitaciones pasa por aprovechar distintas fuentes de variación en la exposición a la

política en cuestión. En el marco de determinados programas de elección escolar, una de estas fuentes puede ser un determinado mecanismo de aleatorización. Es el caso de los estudios que evalúan los impactos de programas de cheques escolares (school vouchers) dirigidos a ampliar el margen de elección de centro de las familias socioeconómicamente más desfavorecidas; cheques que, en un contexto de sobredemanda, acaban siendo distribuidos de forma aleatoria mediante un sorteo. Estos esquemas han sido habituales en diferentes ciudades de los Estados Unidos[30], y han generado un volumen importante de evidencia experimental sobre su efectividad[31].

También existen estudios que han evaluado los efectos del margen abierto a la elección de centro mediante métodos cuasiexperimentales. Es aquí donde encontramos aplicaciones interesantes de la técnica basada en variables instrumentales.

A modo de ejemplo, nos referimos a la evaluación de Gibbons et al. (2008) sobre los impactos que tienen los contextos de elección de centro y competencia entre colegios sobre el rendimiento académico de los alumnos ingleses de primaria. En concreto, se miden los efectos sobre las ganancias competenciales en lengua y matemáticas que experimentan los alumnos entre el comienzo y el final de la educación primaria, tomando como muestra el alumnado de los colegios del área metropolitana de Londres.

En este estudio se define el margen de elección de centro de las familias considerando el número de centros del distrito escolar que las incluyen dentro de un determinado radio de proximidad[30]. De esta manera se comprueba que las familias que residen cerca del límite del distrito disponen de un margen de elección menor que las que viven en zonas más céntricas. Esta circunstancia (concretamente, la distancia entre la residencia de las familias y los límites de sus distritos escolares) se adopta como variable instrumental a los efectos de la evaluación de impacto correspondiente. El supuesto de fondo es que, a igualdad en el nivel de calidad de los colegios, la probabilidad de que una familia escoja un determinado centro disminuirá cuanto más lejos esté de su domicilio. También se asume que las familias que se instalan lejos de los límites del distrito por motivos de escolarización lo hacen, no tanto porque valoren el margen de elección en sí mismo, sino, en todo caso, porque persiguen estar cerca de determinados colegios. En conclusión, la distancia entre residencia y límites del distrito se muestra como una variable vinculada a la probabilidad de encontrarse o no afectado por un mayor margen de elección de centro, y al tiempo no relacionada de forma específica con la probabilidad de éxito educativo de los hijos de las familias que escogen.

A través de este método, los autores demuestran que el incremento del margen de la elección de centro entre las familias no comporta, por sí mismo, ninguna ganancia en el rendimiento académico de sus hijos ni en lengua ni en matemáticas. De hecho, los efectos se muestran más bien negativos, aunque en ningún caso estadísticamente significativos.

EN CATALUÑA...

- De forma general, corresponde a las comunidades autónomas concretar el procedimiento de admisión del alumnado a los centros sufragados con fondos públicos, sobre la base de determinados criterios de elección y asignación escolar establecidos a nivel estatal. Más allá de las comunidades autónomas, los municipios tienen margen para sustanciar algunos de estos procedimientos y criterios; este es, por ejemplo, el caso de la definición de la proximidad a los centros como elemento de priorización en el acceso.
- Esto ha dado lugar en la práctica a la existencia de modelos autonómicos y planificaciones municipales de elección escolar muy distintos y cambientes en el tiempo. Un caso interesante es, por ejemplo, la ciudad de Barcelona, que en el curso 2006-2007 pasó de un modelo de asignación zonal a un modelo de colegios de proximidad. El nuevo modelo trataba de ofrecer a todas las familias un mínimo de seis centros (tres públicos y tres concertados) donde se obtuviera máxima puntuación en el baremo de proximidad. En el curso 2011-2012 este mínimo se amplió a doce centros (seis públicos y seis concertados).
- Toda esta variabilidad entre territorios y en el tiempo responde poco a acciones tomadas sobre la base de evidencias empíricas acerca de su efectividad. Esta variabilidad tampoco se ha aprovechado para generar evidencia, en Cataluña, sobre la capacidad de impacto de las políticas de elección de centro en el terreno del rendimiento y la equidad educativa[33].

4.3 EL USO DE LOS MODELOS DE DOBLES DIFERENCIAS

4.3.1 LA COMPRENSIVIDAD DEL SISTEMA EDUCATIVO

El nivel de comprensividad del sistema escolar, es decir, el tiempo que los alumnos comparten un mismo itinerario formativo antes de iniciarse la separación entre la vía académica y la profesional, es una de las características básicas de su configuración formal. Numerosos estudios han tratado de determinar su incidencia en los logros de los estudiantes recurriendo a pruebas internacionales de competencias (Ammermüller, 2005; Dupriez et al., 2008; Duru-Bellat & Suchaut, 2005a; Gorard & Smith, 2004; Woessmann et al., 2009). Estos estudios suelen evidenciar la presencia de una asociación significativa entre el nivel de comprensividad de los países y el alcance de las desigualdades educativas de sus alumnos: a mayor diferenciación (early tracking), más distancias de rendimiento y más reflejan dichas distancias las desigualdades socioeconómicas y culturales de los estudiantes. Y tampoco parece que los sistemas más diferenciados superen a los más comprensivos en media global de rendimiento.

La presencia de estas asociaciones, sin embargo, no implica necesariamente la existencia de una relación de causalidad entre variables. En efecto, es esperable que el modelo de comprensividad de los países se encuentre correlacionado con otras características no observadas de carácter contextual (nivel de riqueza u homogeneidad cultural del país, características de sus redes escolares, etcétera) que, a su vez, se encuentren relacionadas con los resultados formativos de sus estudiantes.

Demostrar que el early tracking es efectivamente causante de desigualdad educativa y que, por tanto, no origina ganancias en términos de rendimiento general, precisa de un paso más. Los dos ejemplos que mencionamos a continuación dan este paso de la mano de una estrategia de identificación basada en el uso de modelos de dobles diferencias.

El primer ejemplo lo encontramos en el estudio de Hanushek y Woessmann (2006) sobre los efectos del early tracking en los logros y desigualdades académicas de los alumnos en distintos países de la OCDE. La estrategia utilitzada por los autores consiste en comparar las diferencias entre los resultados que los alumnos obtienen en la educación primaria (siempre comprensiva) y los que obtienen en la secundaria (comprensiva en algunos países, diferenciada en otros) en países con y sin early tracking. Este estudio ilustra una aplicación internacional de las dobles diferencias donde la comparación se establece entre conjuntos de países y donde los países son observados no en dos momentos diferentes del tiempo, sino, cada uno de ellos, en dos etapas educativas distintas. La conclusión de esta investigación corrobora los argumentos de los estudios correlacionales antes mencionados, a saber: que el early tracking aumenta las desigualdades sociales de rendimiento sin provocar ningún incremento en la media de resultados de los países.

Un segundo ejemplo de evaluación de los impactos de la comprensividad basada en un modelo de dobles diferencias nos lo brindan aquellos estudios que aprovechan la circunstancia de que esta reforma haya sido desplegada gradualmente en el territorio —regiones dentro de un estado, por ejemplo— según una cronología ajena a factores que podrían afectar a los outcomes seleccionados (Meghir & Palme, 2005; Pekkarinen, Uusitalo, & Kerr, 2009). Este es el enfoque que utilizan Felgueroso, Gutiérrez y Jiménez (2013) para evaluar el impacto de la reforma comprensiva contenida en la LOGSE sobre los niveles de abandono educativo prematuro de los jóvenes. En efecto, después de su aprobación en el año 1990, la reforma fue implantada gradualmente en las distintas comunidades autónomas y lo hizo siguiendo un calendario no determinado por motivos susceptibles de explicar el porqué de unos u otros outcomes finales. Así, a través de un modelo de dobles diferencias, los autores observan cómo, en comparación con el sistema previo que ofrecía la opción de cursar FP-1 a los alumnos de 14 a 16 años con menos disposición hacia las enseñanzas académicas, la extensión de la comprensividad hasta los 16 años representó un incremento del abandono entre los chicos y no entre las chicas.

EN CATALUÑA...

- Cuán comprensivo debe ser el sistema educativo ha sido y sigue siendo una cuestión central en toda reforma educativa de cierto calado. Así, en España, la extensión de la educación comprensiva hasta los 16 años —hecha efectiva con la introducción de la ESO— representó uno de los principales hitos de la LOGSE. De igual forma, la reducción del tramo comprensivo al primer ciclo de la ESO se convierte en uno de los puntos nodulares de la reforma contenida en la Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE, de 9 de diciembre de 2013). En efecto, la LOMCE sitúa el inicio de la diferenciación de itinerarios en 4.º de la ESO (considerado el segundo ciclo de la ESO), momento en el que se distinguen las opciones de enseñanzas académicas de iniciación al Bachillerato, enseñanzas aplicadas de iniciación a la Formación Profesional y los ciclos de Formación Profesional Básica (de dos años de duración comparables con los antiguos Programas de Cualificación Profesional Inicial). Si bien la introducción de la FP Básica es una realidad desde el curso 2014-2015, la diversificación del segundo ciclo de la ESO comenzaría a aplicarse a partir del curso 2016-2017.
- Sin embargo, también en este caso es necesario remarcar el hecho de que ni la reforma comprensiva que desplegó la LOGSE ni la reducción de la comprensividad operada por la LOMCE se han apoyado en evidencias empíricas sólidas sobre la efectividad de unas opciones y otras. En nuestro entorno, este ha sido un debate marcadamente politizado que ha vivido muy de espaldas a la generación y aprovechamiento de conocimiento sobre qué funciona y qué no funciona con relación a esta cuestión.

4.3.2 LA DURACIÓN DE LA JORNADA ESCOLAR

Estudios basados en los resultados de pruebas internacionales estandarizadas acostumbran a detectar una asociación positiva, aunque generalmente débil, entre el número de horas lectivas que ofrecen los colegios y el rendimiento académico que obtienen los estudiantes (OECD, 2007; Hanushek & Woessmann, 2010). Seguimos, sin embargo, manteniendo la cautela necesaria con los resultados de estos estudios. Incluso cuando comparamos colegios dentro de un mismo país, deberíamos pensar que el hecho de que los centros establezcan esquemas horarios distintos no es ajeno a otros aspectos relativos a su organización, sus recursos, su profesorado o a su alumnado, a menudo difíciles de objetivar y, en cambio, previsiblemente relacionados con los resultados educativos de sus estudiantes.

Tratando de superar la incidencia de estos posibles sesgos, algunos estudios han podido analizar la efectividad de medidas de reorganización de los horarios lectivos mediante aproximaciones cuasiexperimentales rigurosas. En este terreno, los modelos de dobles diferencias han tenido un recorrido interesante.

Ponemos como ejemplo el estudio de Lavy (2012) en el Estado de Israel. El autor evalúa los efectos de una reforma introducida en el año 2004 en la educación primaria, la cual altera el sistema de cálculo de la asignación presupuestaria de los centros. A raíz de este cambio los colegios socioeconómicamente más desfavorecidos vieron incrementado su presupuesto, mientras que el de los colegios más favorecidos se redujo; el resto de centros se mantuvieron igual que estaban antes de la reforma. También a consecuencia de esta modificación, los colegios que incrementaron su dotación presupuestaria tendieron a ampliar el número de horas lectivas por semana, en especial las dedicadas a asignaturas troncales.

Aprovechando esta serie de circunstancias, el autor plantea una evaluación de los impactos de la reforma (en particular, del incremento de los recursos de los centros y de las horas lectivas), mediante el uso de modelos de dobles diferencias. En concreto, la evaluación cuenta con una muestra de 936 colegios y la posibilidad de observar sus resultados entre los años 2002 (prerreforma) y 2005 (posrreforma). Un primer tipo de estimaciones se dirige a comparar la evolución de outcomes, antes y después de la reforma, de los colegios que incrementaron sus recursos (grupo de tratamiento) con la evolución de aquellos que no experimentaron cambios en su dotación presupuestaria (grupo de comparación). Un segundo bloque de estimaciones contrasta los resultados de los colegios perjudicados por la reforma (grupo de tratamiento) con los de los colegios no afectados (grupo de comparación). Cada bloque de estimaciones incluye cálculos específicos del efecto de los distintos cambios ocasionados por la reforma: en presupuesto, en horas lectivas por semana y en horas dedicadas a asignaturas troncales. Mediante estos procedimientos, el estudio demuestra que una parte significativa de los beneficios que tiene el incremento de los recursos de los colegios sobre el rendimiento de los alumnos en matemáticas y lengua inglesa son atribuibles a la ampliación del tiempo lectivo dedicado a estas asignaturas.

Los resultados de la evaluación de Lavy concuerdan con las conclusiones de las revisiones de la literatura internacional realizadas por Patall et al. (2010), Redd et al. (2012) y Kidron y Lindsey (2014) sobre esta cuestión. Estos estudios apuntan, además, la especial efectividad de aquellas intervenciones que ocupan el incremento horario con actividades instructivas implementadas por profesores cualificados. Las tres revisiones coinciden en señalar la falta de evaluaciones de impacto rigurosas que caracteriza a este ámbito de intervención.

EN CATALUÑA...

- En el curso 2006-2007, el entonces llamado Departament d'Educació introdujo la denominada sexta hora en la jornada escolar de todos los colegios públicos catalanes. Según argumentaron los responsables de esta medida, y los que la defendieron desde diversos espacios, su principal propósito era el de incrementar las oportunidades educativas de los alumnos de los colegios públicos, resolviendo así el «agravio comparativo» que padecían respecto a los alumnos de los colegios concertados (que tradicionalmente han tenido una jornada lectiva de seis horas diarias). El supuesto que había detrás de este argumento, es decir, que más horas lectivas equivale a más aprendizajes y mejor rendimiento, no vino respaldado por ninguna revisión de evidencias o de análisis ex ante, ni tampoco se contrastó empíricamente en qué medida el agravio horario de la enseñanza pública podía estar repercutiendo sobre el rendimiento de sus alumnos (o sobre otros outcomes no escolares).
- Cinco cursos después de su implantación, el 2011-2012, el Gobierno optó por suprimir la sexta hora en la mayoría de centros, alegando la falta de resultados educativos. En el curso 2013-2014, la sexta hora se mantiene en aproximadamente 400 centros de primaria, colegios situados en entornos especialmente desfavorecidos. En el resto de centros, la supresión de la sexta hora fue sustituida por el recurso SEP (Servicio Escolar Personalizado), dirigido a los alumnos con más dificultades de aprendizaje de cada colegio. De igual modo que la introducción de la medida no respondió a ninguna evaluación de la situación previa, su supresión selectiva tampoco ha sido avalada por ninguna evidencia empírica.

Notas:

22 *Concretamente, los autores utilizan un modelo de regresión para discontinuidad «difusa» (fuzzy). Este modelo se aplica cuando la discontinuidad existente en torno al punto de corte de la variable en cuestión no es nítido (sharp). En el diseño fuzzy, la asignación al tratamiento no es determinístico; así, el impacto del tratamiento se infiere sobre la base de la probabilidad de ser tratado en torno al punto de corte.*

23 *La misma normativa introducía también la repetición automática para los alumnos que contabilizasen más de 25 ausencias diarias no justificadas a lo largo del curso. Esta discontinuidad es aprovechada por Manacorda en el segundo estudio que el autor dirige a la evaluación de impacto de la repetición de curso (Manacorda, 2008).*

24 *Como en el estudio de Jacob y Legren (2009), Manacorda aplica un diseño de regresión discontinua "fuzzy".*

25 *Podríamos imaginar que, a igualdad de otras características, los alumnos y familias más motivados y con una mayor adhesión escolar se esfuerzan más por conseguir una beca que el resto, y que son justamente estos niveles de motivación y adhesión —y no tanto el hecho de haber conseguido una beca— lo que acaba explicando su éxito formativo.*

26 *Los autores estiman que este importe equivale a una tercera parte de los gastos de manutención (alojamiento incluido) de un estudiante universitario a lo largo de un curso académico. En el año 2010, el coste de la matrícula universitaria en Francia se situaba en torno a los 200 euros al año (para grados y másteres).*

27 *Hay que decir, no obstante, que la metodología utilizada en este estudio (basada en la técnica del propensity score matching), si bien equipara a los alumnos tratados (beneficiarios) y a los alumnos control (no beneficiarios) en toda una serie de características observables relevantes, no asegura que unos y otros sean comparables. Podría darse perfectamente que unos y otros difieran en variables no observables que puedan estar relacionadas con el outcome de interés, principalmente, disposiciones individuales y familiares para solicitar una beca y rendimiento académico previo.*

28 *La investigación sobre los efectos del tamaño de la clase cuenta con uno de los experimentos más conocidos en el ámbito de la evaluación de la política educativa: el proyecto STAR (Student-Teacher Achievement Ratio), implementado en el estado de Tennessee (Estados Unidos) entre 1985 y 1989. Sin embargo, la implementación del experimento y, por tanto, la fiabilidad de sus resultados han sido seriamente cuestionados (Hanushek, 1999).*

29 *Incluso cuando comparamos familias dentro de un mismo país, hay que contar con la más que probable incidencia de sesgos entre participantes (familias que escogen de forma activa) y no participantes (familias que no lo hacen). Por mucho que se parezcan en características sociodemográficas relevantes, una familia que ejerce la elección de centro de forma activa seguramente será diferente de otra que la ejerce de forma pasiva en cuestiones que probablemente incidirán en los resultados académicos de sus respectivos hijos.*

30 *Casos paradigmáticos son los programas de cheques de Milwaukee, Cleveland, Nueva York, Dayton o Washington, D.C., todos ellos iniciados durante la década de los años noventa del siglo XX.*

31 *Véanse, por ejemplo, las evaluaciones experimentales de Mayer et al. (2002), Barnard et al. (2003) o Chingos y Peterson (2012) sobre el programa de cheques escolares de la ciudad de Nueva York.*

32 *El peso que los distritos y los colegios otorgan a la proximidad como criterio de priorización del acceso escolar en caso de sobredemanda es muy variable. Como también lo es la manera en que este criterio es interpretado y medido. En todo caso, la definición que Gibbons et al. (2008) ofrecen del radio de proximidad de los centros tiene en cuenta la mediana de las distancias entre la residencia de las familias y los colegios que escogen en cada uno de los distritos.*

33 *En Cataluña deberíamos referirnos a los estudios de Calsamiglia (2013) y Benito y González (2007), el primero sobre la incidencia que tuvo en el año 2007 el cambio del modelo de asignación escolar de Barcelona sobre patrones de elección de colegios de las familias; el segundo sobre la influencia de la zonificación escolar en los niveles de segregación escolar de distintos municipios catalanes. De esta manera, ambas investigaciones estudian los efectos de variaciones regulativas sobre outcomes que podríamos considerar de naturaleza intermedia.*

5. Y EN CATALUÑA, ¿CÓMO PODRÍAMOS AVANZAR?

La preocupación por la efectividad de las intervenciones educativas tiene su origen en la apuesta por promover decisiones y políticas educativas bien informadas, es decir, basadas en la evidencia empírica. La política educativa catalana, como la española y la de tantos otros países de nuestro entorno, está todavía lejos de esta situación. En efecto, según hemos tenido ocasión de ejemplificar, las decisiones sobre las políticas educativas —sobre su lanzamiento, su mantenimiento, su reforma, su supresión— raramente se han basado en la generación o disposición de evidencias empíricas sólidas sobre los impactos de unas u otras intervenciones.

Partiendo de lo que podemos extraer de estudios y evaluaciones realizadas en otros contextos, los argumentos que se plantean a continuación quieren situar algunos espacios de oportunidad para impulsar la práctica de la evaluación de impacto de las políticas educativas en Cataluña.

5.1 ATREVERSE A EXPERIMENTAR: APROVECHAR LAS PRUEBAS PILOTO...

Como hemos indicado anteriormente, no hay diseño de evaluación más robusto que el experimental. Y si bien es cierto que la utilización de este diseño no siempre es viable, también lo es que son numerosas sus posibles aplicaciones. El repertorio de ejemplos que hemos descrito en el capítulo 3 lo pone de manifiesto.

El tipo de problemáticas y los ámbitos de intervención que repasábamos entonces ciertamente conectan con retos importantes que la política educativa catalana tiene planteados actualmente. Algunos de estos ámbitos tienen que ver con elementos estructurales o contextuales de la política y el sistema educativo (autonomía escolar, escolarización en la primera infancia o los incentivos económicos al profesorado); otros remiten a prácticas más «micro», de innovación o desarrollo pedagógico (el uso de las TIC en la enseñanza, las tutorías individualizadas o los programas de implicación educativa de las familias).

En Cataluña no son pocos los programas educativos que se han iniciado en forma de prueba piloto: el proyecto de autonomía de centros (desde 2005-2006), el programa de apoyo escolar personalizado (iniciado el curso 2011-2012), la introducción de la compactación horaria en la ESO y en primaria (desde 2012-2013), el programa intensivo de mejora de la ESO (desde 2013-2014), la modalidad blended learning en el bachillerato (desde 2013-2014), el programa mSchools (desde 2013-2014), el plan experimental de plurilingüismo (desde 2013-2014), la asignatura de servicios a la comunidad (desde 2014-2015), entre otros. En general, el objetivo principal de estas experiencias piloto se ha orientado, por encima de todo, a probar determinados aspectos relativos al diseño y la implementación de la intervención, contrastar su viabilidad técnica y política, su aceptación social; todo ello antes de proceder a su eventual generalización. En cambio, rara vez estas pruebas piloto han partido de una estrategia evaluativa bien definida y dirigida a generar las condiciones que deberían permitir

la estimación posterior de su impacto. El camino a recorrer en este terreno es, por tanto, largo, aunque también está lleno de oportunidades.

Por ejemplo, parece claro que el ámbito de las intervenciones que designamos como más «micro» (innovaciones pedagógicas, estrategias docentes, modificaciones curriculares, incluso iniciativas educativas extraescolares) ofrece un campo particularmente fértil al desarrollo de pruebas piloto de impacto, en particular, de programas piloto experimentales. La misma dinámica de la innovación estimula el ensayo y la experimentación, una dinámica que podría articularse mediante pruebas aleatorizadas a pequeña o mediana escala. De hecho, en contextos en los que todo el mundo tiene claro que se está «probando» una iniciativa educativa, de alcance limitado y con una capacidad de impacto todavía por determinar, suele aceptarse que habrá candidatos interesados (colegios, alumnos, familias, territorios) que no accederán al programa y que el criterio más justo y operativo para asignar la participación pasa por la aleatorización.

Y es también en estos contextos donde la utilidad de la evaluación se observa más claramente. Por un lado, ensayo experimental y recopilación de evidencias se combinan para intentar aportar un conocimiento sólido y útil para el conjunto del profesorado y los responsables de los programas de innovación educativa. Por otro lado, es la obtención de este mismo conocimiento el que permite fundamentar la conveniencia (o no) de escalar los programas piloto en un sentido u otro.

Eventualmente, la iniciativa de estas pruebas piloto puede nacer del territorio y la comunidad educativa (municipios o conjunto de municipios, redes de colegios, de entidades educativas, movimientos y federaciones de profesores, etc.), y disponer para su desarrollo del apoyo y financiación de fundaciones privadas u organizaciones filantrópicas. En todo caso, entendemos que estas iniciativas deberían contar con el apoyo del Departament d'Ensenyament, principal responsable del posible escalado de los programas en cuestión.

El rol del Departament sería preeminente cuando se piloten aspectos estructurales de la política o el sistema educativo. Seguramente, los desafíos y problemas de viabilidad técnica y política que las pruebas piloto experimentales pueden plantear en estos ámbitos de actuación más nucleares son superiores a los que plantean en el terreno de los programas de carácter más «micro». Esto no quiere decir que las políticas de autonomía escolar, de elección de centro, de profesorado, de organización de la jornada escolar, de planificación de la educación infantil, de regulación de la repetición o promoción de curso, etc. no admitan la posibilidad de pilotos experimentales. Pensaríamos entonces en la posibilidad de realizar pilotos orientados a medir el impacto de unos componentes u otros, aplicaciones o reformas concretas que puedan plantearse o formen parte de cada una de estas políticas.

Por ejemplo, podría valorarse la posibilidad de testar mediante pruebas piloto experimentales el despliegue de aspectos específicos de la autonomía escolar, como puedan ser determinados

esquemas de incentivación económica del profesorado o determinados elementos de profesionalización de los equipos directivos. O podrían diseñarse pilotos para evaluar la efectividad de distintos modelos de provisión de la educación de cero a tres años o distintas modalidades de ayudas a las familias para acceder a estos recursos. O pilotar de forma experimental diferentes fórmulas de agrupación de alumnos en diversas asignaturas o cursos. De igual forma, se habrían podido desarrollar auténticas pruebas piloto, con una muestra suficiente de centros y con un mecanismo de aleatorización de la participación, dirigidas a establecer con rigor los efectos de la jornada compacta en la educación secundaria y en la primaria sobre las oportunidades educativas de los alumnos. En la tabla 1 se ofrecen más ejemplos.

5.2 ... Y EL EXCESO DE DEMANDA

El exceso de demanda de un programa puede estar «inducido» en tanto en cuanto la limitación de su cobertura es deliberada; un caso típico de esta situación es el de los programas piloto. Por el contrario, en el ámbito educativo, es habitual encontrar servicios o programas con un exceso de demanda «real»; es decir, los límites en la cobertura obedecen a la incapacidad objetiva del servicio o programa de atender a todo el que lo demanda.

Pues bien, cuando en determinados ámbitos de intervención no se considera viable el planteamiento de pruebas piloto aleatorizadas, pero sí existen programas sobredemandados, sería necesario preservar la posibilidad de aprovechar esta circunstancia a los efectos de evaluar experimentalmente sus impactos. Como dijimos en su momento, en situación de escasez de recursos, cuando el programa no puede cubrir a toda la población a la que se dirige (alumnos, centros, familias con unas características u otras), la asignación aleatoria del recurso entre los solicitantes elegibles es el mecanismo de asignación más justo que podamos diseñar, y el que ofrece las mejores condiciones para la evaluación de impacto.

De hecho, el acceso a determinados servicios educativos con exceso de demanda ya se dirime actualmente mediante un sorteo. Un claro ejemplo es el del acceso a las guarderías públicas (escoles bressol), un recurso a menudo sobredemandado en muchos municipios de Cataluña. En este caso, por ejemplo, se trataría de acabar comparando el progreso educativo de los niños que, en unos municipios y otros, «ganan» la lotería de acceso a la guardería (o a aquellas guarderías con más sobredemanda) con el progreso de los niños que, con los mismos puntos en la solicitud de preinscripción, «pierden» el sorteo (controlando por las opciones educativas alternativas escogidas por las familias de estos niños). Algo similar podría llegar a plantearse para evaluar los impactos de la elección de centro en la educación obligatoria: se trataría de comparar los resultados académicos de los alumnos que entran en los colegios sobredemandados escogidos en primera opción con los de los alumnos que, con los mismos puntos en la baremación, priorizan, sin éxito, los mismos colegios.

Un procedimiento similar permitiría evaluar de forma rigurosa aquellos programas y servicios educativos extraescolares que llegan a encontrarse sobredemandados: actividades de refuerzo escolar, dispositivos de acompañamiento educativo, programas de ocio educativo, etc. Es habitual que las entidades proveedoras de estos servicios apliquen criterios de selección de candidatos que van más allá de los requisitos de elegibilidad establecidos formalmente, criterios que pueden recurrir a juicios sobre el nivel de motivación de los individuos, capacidad para aprovechar el recurso, grado de vulnerabilidad, etc. Incluso en estas circunstancias, si se consiguen baremar estos criterios, podría existir todavía margen para la introducción de mecanismos de sorteo entre individuos empatados en puntos allí donde se produzca sobredemanda.

En cualquier caso, es necesario tener previstos los sistemas de información necesarios para registrar al conjunto de candidatos que demandan el recurso, programa o servicio, ya superen o no el número total de plazas disponibles, y establecer mecanismos que hagan posible la obtención de datos y seguimiento de outcomes de todos los candidatos. Retomaremos esta cuestión en el punto 5.5.

5.3 EVALUACIÓN CUASIEXPERIMENTAL

No infravaloramos los problemas de viabilidad (técnica, política, social) que los estudios experimentales pueden llegar a plantear, sobre todo cuando se discute la evaluación de políticas o regulaciones educativas de índole más sistémica. Cuando el diseño experimental no es viable, es decir, cuando el acceso al programa no puede dirimirse por la vía de la aleatorización, entonces pueden entrar en juego distintas alternativas de evaluación cuasiexperimental.

Más adelante hemos recogido ejemplos de evaluaciones cuasiexperimentales que consiguen producir estimaciones robustas de impacto en ámbitos de intervención que son también relevantes en Cataluña y que podrían ser evaluados aquí siguiendo aproximaciones similares. Por ejemplo, podría valorarse la posibilidad de aplicar diseños de regresión discontinua en la evaluación de programas de becas financiadas y/o gestionadas por la Generalitat de Cataluña que definen la elegibilidad a partir de un punto de corte en una variable, ya sea la renta (becas de comedor, becas en secundaria posobligatoria o becas Equitat en la universidad) o el rendimiento académico (becas especiales en secundaria o becas Excel•lència en la universidad). Este mismo método podría utilizarse para estimar los impactos de repetir curso, tanto si acaba instaurándose la reválida externa que plantea la LOMCE en el último curso de la ESO (cuyos resultados serían vinculantes a los efectos de graduarse o repetir curso), como en el marco actual, en el que la repetición o la promoción de curso se establece de acuerdo al número de materias suspendidas en la evaluación final de curso.

O bien podría explorarse la aplicabilidad de variables instrumentales relacionadas con la distribución o densidad territorial de la red de colegios en la evaluación de determinados cambios en la regulación de la elección escolar, principalmente en municipios grandes o medianos. En otro ámbito, si nos preocupan los efectos que la ratio de alumnos por aula puede tener sobre sus aprendizajes, podríamos llegar a testar las fluctuaciones en el momento del nacimiento como instrumento del tamaño de las clases en los distintos cursos escolares. Este mismo instrumento podría utilizarse para evaluar reformas que conduzcan a incrementar de forma general la ratio máxima de ocupación de las aulas. En nuestro caso se trataría de comparar alumnos en grupos-clase afectados y no afectados por la reforma al inicio de la escolarización, asumiendo que los grupos afectados alcanzan la ratio máxima por aula y que los grupos no afectados no incrementan la ratio en cursos sucesivos.

Finalmente, podría estudiarse la aplicación de modelos de dobles diferencias para evaluar la efectividad de programas o medidas que han tenido un despliegue gradual en el territorio o en la red de centros educativos. Este fue el caso de la jornada lectiva de seis horas en educación primaria. Recordemos que la denominada sexta hora fue introducida en el curso 2006-2007 por el entonces Departament d'Educació en todos los centros públicos de primaria; en el curso 2011-2012, esta medida fue retirada en un número significativo de colegios. Este hecho abriría la puerta a analizar en qué medida las diferencias de rendimiento que, a partir del curso 2011-2012, se observan entre centros con y sin sexta hora se corresponden con las diferencias que estos mismos grupos de centros mostraban en el período en que todos ellos disponían de este recurso. La comparación entre estas dos diferencias nos acercaría a la medida del impacto de la sexta hora.

5.4 ... PREVIENDO LA EVALUACIÓN EN EL DISEÑO DEL PROGRAMA

La posibilidad de optar por las distintas alternativas de evaluación cuasiexperimental (regresiones discontinuas, variables instrumentales, dobles diferencias u otros métodos no considerados aquí) dependerá del momento en que estas se consideren. Se dibujan aquí dos escenarios: cuando el diseño de la política contempla el diseño de la evaluación (escenario ex ante) o cuando la evaluación se diseña una vez que el programa ya está en marcha o ha finalizado (escenario ex post).

En el primero de los escenarios, la propia formulación de la intervención —definición de objetivos, población destinataria, fases de ejecución, actividades, recursos, etc. — incorpora la preocupación por sus condiciones de evaluabilidad. Hablamos entonces de perspectiva ex ante o prospectiva de la evaluación de impacto. Esta perspectiva, que es inherente a los experimentos sociales, incrementa la probabilidad de acabar obteniendo buenas evaluaciones de impacto, también en ausencia de aleatorización, principalmente por dos motivos: por un lado, permite prever fórmulas de implementación que faciliten la construcción de buenos grupos de comparación; por otro lado, ofrece la posibilidad de establecer desde el inicio

los instrumentos y las medidas que requerirá la evaluación del programa (indicadores de resultado, sistemas de información, protocolos de recogida de información, etc.). Más allá de todo esto, como indican Casado y Todeschini (2013), el simple hecho de pensar cómo evaluar el impacto del programa en el mismo momento en que se está diseñando, «obliga a los responsables del programa a precisar nítidamente cuáles son los outcomes de interés sobre los que la intervención pretende tener algún impacto, así como los mecanismos por los que se espera que estos efectos se produzcan» (2013: 22). Todo ello redunda no solamente en una mejora de la evaluabilidad del programa, sino también del diseño sustantivo de la intervención.

En el segundo escenario, cuando la evaluación se diseña ex post (una vez el programa se encuentra en marcha o ya ha finalizado), su evaluabilidad se encuentra totalmente condicionada a las características de la implementación del programa y a las bases de información disponibles. En el peor de los casos, puede llegar a resultar imposible identificar grupos de comparación mínimamente creíbles u obtener la información necesaria para el seguimiento de los resultados en cuestión.

Por tanto, deberá apostarse claramente por propiciar el primero de los escenarios siempre que sea posible. Hay que decir que esta apuesta se deja notar en una parte importante de las convocatorias para programas educativos de instituciones internacionales como el Banco Iberoamericano de Desarrollo o la propia Comisión Europea, los cuales supeditan la concesión de financiación al diseño ex ante de evaluaciones de impacto.

Tabla 2. Oportunidades para la evaluación de impacto de políticas y programas educativos en Cataluña

ÁMBITO DE INTERVENCIÓN	OPORTUNIDADES NO APROVECHADAS	ESPACIOS DE OPORTUNIDAD
SISTEMA EDUCATIVO		
Comprensividad frente a diferenciación temprana	• Reforma comprensiva, LOGSE (G.E., 1990): no fundamentada empíricamente (ex ante), evaluación retrospectiva limitada (ex post) • Reforma de la secundaria, LOMCE (G.E., 2013): no fundamentada empíricamente (ex ante)	• Posibilidad de explotar la introducción gradual de la reforma comprensiva en el territorio como estrategia de identificación de contextos «tratados» y «control» (modelos dobles diferencias)

Educación en la primera infancia	• Reforma comprensiva LOGSE (G.E., 1990): no fundamentada empíricamente (ex ante), evaluación retrospectiva limitada (ex post) • Expansión de plazas de guardería (G.C. y G.L., entre 2005 y 2010): no fundamentada empíricamente (ex ante), evaluación retrospectiva limitada (ex post) • Reducción aportaciones a guarderías (G.C., desde 2010): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de explotar el mecanismo de aleatorización ante el exceso de demanda en el acceso a guarderías (diseño experimental) • Posibilidad de explotar la introducción gradual de la reforma comprensiva en el territorio como estrategia de identificación de contextos «tratados» y «control» (modelos dobles diferencias)
CUASIMERCADO EDUCATIVO		
Elección de centro	• Mecanismos de acceso/asignación escolar (G.C. y G.L., años diversos): no fundamentados empíricamente (ex ante), evaluación retrospectiva limitada (ex post)	• Posibilidad de explotar el mecanismo de aleatorización ante el exceso de demanda en el acceso a determinados colegios en determinados entornos (diseño experimental) • Posibilidad de aprovechar desequilibrios en la distribución territorial de la red escolar (variables instrumentales)
Autonomía escolar	• Plan de Autonomía de los Centros (G.C., 2005): prueba piloto no aleatorizada, sin evaluación de impacto (ex post) • Decreto de Autonomía de los Centros (G.C., 2010): no fundamentado empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de desarrollar pruebas piloto para testar la efectividad de determinadas medidas de autonomía contempladas en el decreto de 2010 (diseño experimental)
INCENTIVOS Y AYUDAS ECONÓMICAS		
Retribución al profesorado	• Cambios en estructuras retributivas, promoción e incentivos (G.E. y G.C., años diversos): no fundamentados empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de desarrollar pruebas piloto para testar la efectividad de determinados esquemas de incentivos monetarios o no monetarios (diseño experimental)
Becas y ayudas a los estudiantes	• Introducción y cambios en becas escolares (G.E. y G.C., años diversos): no fundamentados empíricamente (ex ante), evaluación retrospectiva limitada (ex post) • Reducción becas escolares (G.E. y G.C., desde 2010): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post) • Programa de becas universitarias: Equitat y Excel•lència (G.C., desde 2011): no fundamentado empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de explotar la variabilidad territorial en la cobertura de becas y en la introducción de cambios en dotaciones como estrategia de identificación de alumnos «tratados» y «control» (variables instrumentales) • Posibilidad de aprovechar discontinuidades en la percepción en torno al punto de corte en el nivel de renta o académico establecido para acceder a determinadas becas (regresiones discontinuas)

ASPECTOS ORGANIZATIVOS		
Alumnos por aula	• Establecimiento de las ratios máximas alumnos por clase, primaria y secundaria (G.E. y G.C., años diversos): no fundamentado empíricamente (ex ante), sin evaluación de impacto (ex post) • Ampliación ratios máximas alumnos por clase, primaria y secundaria (G.E. y G.C., 2012-2013): no fundamentada empíricamente (ex ante)	• Posibilidad de explotar variaciones en las ratios motivadas por fluctuaciones en el tamaño de las cohortes de alumnos y/o por medidas de incremento de ratios (variables instrumentales)
Horarios y jornada escolar	• Introducción 6.ª hora en la primaria pública (G.C., 2006-2007): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post) • Suspensión 6.ª hora en mayoría de centros (G.C., 2011-2012): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post) • Compactación jornada secundaria pública (G.C., 2012-2013): fundamentada en prueba piloto no aleatorizada • Introducción «semana blanca» (G.C., 2011-2012): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post) • Supresión «semana blanca» (G.C., 2012-2013): no fundamentada empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de desarrollar pruebas piloto para testar la efectividad de medidas de ampliación o compactación horaria o cambios en el calendario escolar (diseño experimental) • Posibilidad de aprovechar discontinuidades en la adhesión a la reforma en torno al punto de corte en el nivel de consenso mínimo exigido por parte del consejo escolar (regresiones discontinuas) • Posibilidad de aprovechar la introducción (y retirada) parcial de cambios horarios en la red escolar como estrategia de identificación de contextos «tratados» y «control» (modelos dobles diferencias)
DECISIONES Y PRÁCTICAS PEDAGÓGICAS		
Repetición de curso	• Cambios en modelos repetición-promoción automática (G.E. y G.C. y centros, años diversos): no fundamentados empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de aprovechar discontinuidades en la probabilidad de repetir en torno al número mínimo de asignaturas aprobadas (o al umbral mínimo de puntuación en eventuales reválidas vinculantes) requerido para pasar de curso (regresiones discontinuas) • Posibilidad de comparar alumnos con el mismo rendimiento académico en cursos consecutivos, unos afectados y otros no por cambios en los criterios de repetición (variables instrumentales)

Innovación y desarrollo pedagógico	• Introducción y cambios en programas diversos (TIC, enfoque competencias, sistemas de tutoría, aprendizaje cooperativo...) (G.C. y centros, años diversos): no fundamentados empíricamente (ex ante), sin evaluación de impacto (ex post)	• Posibilidad de desarrollar pruebas piloto para testar la efectividad de determinadas innovaciones curriculares y didácticas (diseño experimental)
INICIATIVAS EXTRAESCOLARES		
Fuera del colegio ***Relación familia-colegio***	• Introducción y cambios en programas diversos (planes de entorno, proyectos comunitarios, educación no formal...) (G.C., G.L. y centros, años diversos): no fundamentados empíricamente (ex ante), evaluación retrospectiva limitada (ex post)	• Posibilidad de desarrollar pruebas piloto para testar la efectividad de programas de implicación de las familias en la escolarización de sus hijos, así como de programas fuera del colegio (diseño experimental) • Posibilidad de aprovechar discontinuidades en la participación en torno al punto de corte en el nivel de renta o académico establecido para acceder a determinados recursos fuera del colegio (regresiones discontinuas)

G.E. = Gobierno español; G.C. = Gobierno catalán; G.L. = Gobierno local

Fuente: *elaboración propia*

5.5 SISTEMAS DE INFORMACIÓN Y ACCESO A LOS DATOS

Cualquier evaluación de impacto que aspire a ser rigurosa requiere buenos datos en los que basarse para construir sus variables independientes (características de los alumnos y sus familias, sus clases, profesores, colegios y entornos, etc.) y sus variables de outcome (indicadores de rendimiento, de trayectoria educativa, o variables relacionadas con expectativas, actitudes o disposiciones escolares). Difícilmente puede avanzarse en el estudio de la efectividad de las políticas y programas educativos cuando: 1) los registros educativos son incompletos (falta de información relevante) o inconexos (sin vinculación entre ellos ni con otros registros administrativos), y cuando: 2) el acceso a la información se encuentra limitado o restringido. En Cataluña hay mucho camino por recorrer en ambos aspectos.

En primer lugar, los registros carecen de información relevante sobre los alumnos, principalmente, sobre el nivel socioeconómico y educativo de sus hogares. Esta información podría agregarse a nivel de colegio y proporcionar indicadores del perfil social de los centros más precisos que los que se utilizan actualmente. Otras informaciones están muy poco sistematizadas, poco tratadas o dispersas; sucede, por ejemplo, con la información sobre las becas escolares (incluidas las de comedor) o con la relativa al perfil y condiciones del profesorado de los centros. Por otra parte, los registros educativos existentes suelen estar poco interconectados, de manera que se dificulta la posibilidad de establecer vinculaciones entre variables de interés; por ejemplo, entre las calificaciones académicas de los alumnos

(ordinarias y en pruebas de competencias) y el hecho de que sean o hayan sido beneficiarios de una beca escolar o de alguna adaptación curricular, o hayan pasado por una guardería, o hayan tenido unos profesores determinados u otros, etc. Y también resulta difícil conectar los registros educativos disponibles con otros registros administrativos como, por ejemplo, el de la Seguridad Social (para conocer la trayectoria y situación laboral de los padres y también de los propios alumnos una vez finalizados los estudios) o el de la Agencia Tributaria (para conocer la situación y evolución de los niveles de renta de las familias y los alumnos en cuestión).

La construcción de un auténtico sistema de registros individuales debería permitir completar e integrar este conjunto de informaciones en torno a un nexo central, el alumno, y a partir de aquí poder proceder a las agregaciones y desagregaciones de datos que requiera cada análisis (por clase, por profesor, por centro o red escolar, barrio, etc.). De esta manera se facilitaría la posibilidad de trazar la caracterización del estudiante a lo largo de toda su trayectoria educativa (incluso más allá).

A falta de este sistema, determinados tipos de evaluaciones de impacto podrían cubrirse con datos de panel, susceptibles de introducir un seguimiento longitudinal de resultados con suficiente recorrido en el tiempo, y siempre y cuando estos paneles incluyan información completa y relevante sobre las variables de interés. Esta herramienta resulta útil para, por ejemplo, recoger y ampliar información sobre outcomes no cognitivos, pero que están vinculados con el progreso académico (disposiciones y expectativas hacia la educación, autoestima, autonomía, capacidad crítica, etc.).

En segundo lugar, buena parte de la información que incluyen actualmente los registros educativos no es de fácil acceso para los investigadores. En efecto, los grupos de investigación a menudo encuentran impedimentos para poder disponer de datos desagregados (anonimizados y disociados[34]) sobre los alumnos o los centros a los que asisten, tanto sobre outcomes educativos (por ejemplo, resultados en evaluaciones ordinarias o en pruebas externas i de rendimiento) como sobre determinades característiques sociodemogràfiques.

En resumen, si queremos llegar a disponer de evidencia sólida sobre el grado de efectividad de programas y políticas educativas, lo primero que debemos conseguir es mejorar sus condiciones de evaluabilidad, entre otras, aquellas que conciernen a los sistemas, fuentes y tipos de información que debieran entrar en juego.

5.6 QUÉ FUNCIONA Y POR QUÉ FUNCIONA: LA IMPORTANCIA DE HILAR FINO

Simplificando mucho, diríamos que las evaluaciones de efectividad de los programas pueden proporcionar dos tipos de evidencias: evidencia de que el programa funciona o evidencia de que el programa no funciona (tiene impactos negativos o no los tiene). Sin embargo, diciendo esto nos quedamos cortos. Veamos por qué.

Primero, es habitual que los programas educativos se fijen más de un objetivo estratégico. Al menos suelen ser diversos los outcomes sobre los que poder valorar sus efectos. Así, es frecuente que un programa consiga impactar positivamente sobre parte de sus outcomes y no sobre otra parte. Nos encontraríamos, por tanto, ante un caso de programa con «efectos mixtos», un matiz que, bastante a menudo, queda eclipsado en el debate público por las categorías generales de «funciona» o «no funciona». Más aún, también es posible que una intervención que no funciona para el conjunto de outcomes identificados, sí resulte positiva en relación con otros outcomes que no han sido tenidos en cuenta y que, sin embargo, están vinculados al propósito final del programa. Obviamente, desde el punto de vista de la opción política que enmarca la intervención, no todos los resultados son igualmente relevantes: algunos tienen que ver con objetivos prioritarios y otros no. Por tanto, no todos los impactos tienen el mismo «valor». Y, no obstante, el hecho de poder diferenciar los efectos logrados sobre unos outcomes y otros aporta un conocimiento de gran importancia de cara a posibles reorientaciones de la intervención en cuestión.

Segundo, las intervenciones educativas, por específicas y focalizadas que sean, acostumbran a incluir actuaciones diversas, articuladas como componentes de una misma estrategia y diseñadas sobre la base de unos objetivos finales comunes. Por tanto, puede perfectamente ocurrir que un programa no funcione y, en cambio, sí funcione alguno de sus componentes particulares, o también a la inversa, que un programa que globalmente funciona contenga componentes que son de hecho inefectivos (o incluso perjudiciales).

Tercero, las intervenciones educativas, de nuevo, por específicas y focalizadas que sean, se dirigen a una población que no deja de presentar características diversas. Esto quiere decir que puede ocurrir que un programa no funcione a la vista de sus resultados medios sobre el conjunto de la población atendida, pero sí resulte positivo cuando se presta atención a lo que les sucede a unos colectivos determinados. Igualmente a la inversa: un programa que es en conjunto efectivo puede estar concentrando sus impactos positivos en unos subgrupos concretos y en cambio resultar inocuo (incluso negativo) para otros subgrupos.

Cuarto, la gran mayoría de programas educativos necesitan su tiempo, en dos sentidos. En el terreno de la implementación, las intervenciones requieren lo que podríamos denominar un tiempo de «rodaje» o «estabilización» antes de que podamos considerar que su desarrollo es medianamente fiel a su diseño inicial (es decir, a lo que se pretende evaluar) y haya podido acomodarse a las lógicas y rutinas particulares de los agentes implicados en cada caso. Esto resulta especialmente cierto en el caso de aquellas intervenciones que representan cambios significativos en las dinámicas y en las prácticas de estos agentes. Diríamos así que los programas requieren tiempo antes de que resulte razonable empezar a observar sus posibles impactos. Al margen de que la intervención pueda definir objetivos mensurables a corto, medio o largo plazo, y al margen del tiempo de rodaje que pueda requerir su implementación, determinados cambios en la realidad tratada llegan con el tiempo. Esta última posibilidad tiene

su reverso: programas que se muestran efectivos a corto plazo, pero con impactos que se disipan con el paso del tiempo (fade-out effects).

Trasladamos estos cuatro argumentos a nuestro ejemplo imaginario, el Programa de Aceleración Educativa (PAE). Supongamos que una evaluación rigurosa de su efectividad acaba concluyendo que el programa, en conjunto, no funciona, no produce impactos significativos. ¿Respaldaría esta constatación una eventual decisión de retirada total del programa? La respuesta sería: «No necesariamente».

En primer lugar, podría ocurrir que el PAE no consiguiera impactos positivos sobre el nivel competencial en matemáticas, pero en cambio sí comportase pequeñas ganancias en el ámbito lingüístico o en outcomes no observados (por ejemplo, mejoras actitudinales o en el terreno socioemocional). En segundo lugar, una evaluación del impacto diferencial de cada uno de sus componentes podría detectar que lo que hace que el programa, en conjunto, no funcione es el ingrediente de agrupación por nivel que incorpora; en cambio, podríamos observar que el incremento del tiempo lectivo dedicado a las asignaturas instrumentales sí muestra impactos positivos significativos en el progreso académico de los alumnos. En tercer lugar, podría ser que el PAE resultara inefectivo al fijarse en la totalidad del alumnado atendido, pero en cambio sí tuviera un impacto específico en el caso concreto de aquellos participantes procedentes de familias con un menor capital socioeconómico y educativo. Finalmente, si pudiéramos disponer de una ventana de observación lo suficientemente amplia, podría ocurrir que detectásemos que el programa no es efectivo durante sus dos primeros años de funcionamiento y que, en cambio, empezara a generar impactos positivos a partir del tercer año. Más aún, podría observarse que determinadas ganancias competenciales solo afloran a medio plazo, cuando el alumno está finalizando la ESO, ganancias que podrían interpretarse como resultado indirecto de otras mejoras en el terreno actitudinal y emocional.

En síntesis, es importante completar la conclusión sobre la efectividad global de un programa con evidencias sobre el detalle de los impactos: saber qué componentes del programa funcionan y cuáles no, averiguar para quién funcionan y para quién no, a partir de qué momento y durante cuánto tiempo. Llegar a este nivel de detalle no siempre resulta fácil. No siempre se dispone de grupos de comparación específicos para los distintos niveles de participación en las distintas fases o componentes del programa; no siempre es sencillo discernir sobre qué subgrupos de participantes tiene sentido analizar la ocurrencia de posibles efectos heterogéneos; no siempre es posible comenzar el proceso de evaluación cuando más conviene (cuando el programa ya ha sido rodado) ni disponer de una ventana temporal lo bastante amplia para la observación de los resultados considerados. Sin embargo, si queremos desentrañar los mecanismos de efectividad (o inefectividad) de las intervenciones educativas y saber qué hace que un programa funcione o no, hay que impulsar evaluaciones de impacto que apunten en esta dirección.

Notas:

[34] *Huelga decir que deberá evitarse siempre el riesgo de incumplir las normas básicas referentes al secreto estadístico.*

REFERENCIAS

Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T. J., i Pathak, P. (2009). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. National Bureau of Economic Research, NBER Working Paper No. 15549. Recuperat a partir de http://www.nber.org/papers/w15549

Alegre, M. A., i Benito, R. (2014). Youth Education Attainment and Participation in Europe: the role of contextual factors and the scope of education policy. European Journal of Education, 49(1), 127-143.

Allen, C. S., Chen, Q., Willson, V. L., i Hughes, J. N. (2009). Quality of Research Design Moderates Effects of Grade Retention on Achievement: A Meta-Analytic, Multilevel Analysis. Educational Evaluation and Policy Analysis, 31(4), 480-499.

Ammermüller, A. (2005). Educational opportunities and the role of institutions. ZEW Discussion Paper No. 05-44. Recuperat a partir de http://www.econstor.eu/handle/10419/24135

Angrist, J., Cohodes, S.R., Dynarski, S.M., Pathak, P.A., i Walters, C.D. (2013). Charter Schools and the Road to College Readiness: The Effects on College Preparation, Attendance and Choice. The Boston Foundation and NewSchools Venture Fund. Recuperat a partir de https://www.tbf.org/~/media/TBFOrg/Files/Reports/Charters%20and%20College%20Readiness%202013.pdf

Angrist, J. D., i Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics, 114(2), 533-575.

Ball, S. J. (2012). Global Education Inc.: New Policy Networks and the Neoliberal Imaginary. London: Routledge, Taylor & Francis Group.

Barnard, J., Frangakis, C. E., Hill, J. L., i Rubin, D. B. (2003). Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City. Journal of the American Statistical Association, K(462), 299-323.

Barnett, W. S. (1985). Benefit-cost analysis of the Perry Preschool Program and its policy implications. Educational evaluation and policy analysis, 7(4), 333-342.

Belfield, C. R., Nores, M., Barnett, S., i Schweinhart, L. (2006). The High/Scope Perry Preschool Program Cost–Benefit Analysis Using Data from the Age-40 Followup. Journal of Human Resources, 41(1), 162-190.

Benito, R., i Gonzàlez, I. (2007). Processos de segregació escolar a Catalunya. Barcelona: Mediterrània, Col. Polítiques 59.

Bettinger, E. P., Long, B. T., Oreopoulos, P., i Sanbonmatsu, L. (2012). The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment. The Quarterly Journal of Economics, 127(3), 1205-1242.

Betts, J. R., i Tang, Y. E. (2011). The Effect of Charter Schools on Student Achievement. A Meta-analysis of the Literature. University of Washington and Center on Reinventing Public Education. Recuperat a partir de http://www.oxydiane.net/IMG/pdf/Charter_NCSRP_BettsTang_Oct11.pdf

Blasco, J. (2015). Infància a Catalunya. Mesures contra la pobresa infantil: ampliació selectiva d'escoles bressol i extensió de la tarifació social. Barcelona: UNICEF, Docs Infància a Catalunya. Recuperat a partir de http://www.unicef.es/sites/www.unicef.es/files/docs_infancia_cat_escoles_bressol_.pdf

Blasco, J., i Casado, D. (2009). Guia pràctica 5. Avaluació d'impacte. Ivàlua. Recuperat a partir de http://www.ivalua.cat/documents/1/01_03_2010_11_33_12_Guia5_Impacte_Setembre2009_revfeb2010_massavermella.pdf

Bonal, X., i Verger, A. (2013). L'Agenda de la política educativa a Catalunya: una anàlisi de les opcions de govern (2011-2013). Barcelona: Fundació Jaume Bofill, Col. Informes Breus 45.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., i Chambers, B. (2007). Final Reading Outcomes of the National Randomized Field Trial of Success for All. American Educational Research Journal, 44(3), 701-731.

Burch, P. (2009). Hidden markets: The new education privatization. New York: Routledge.

Calsamiglia, C., i Güell, M. (2014). The Illusion of School Choice: Empirical Evidence from Barcelona. IZA DP No. 8202. Recuperat a partir de http://ftp.iza.org/dp8202.pdf

Camilli, G., Vargas, S., Ryan, S., i Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. The Teachers College Record, 112(3), 579-620.

Carter, S. C. (2000). No Excuses: Lessons from 21 High-Performing, High-Poverty Schools. Washington, D.C.: The Heritage Foundation.

Casado, D. (2012). Per què no avaluem les polítiques públiques com els fàrmacs? Una aposta per l'experimentació social. Avaluació per al Bon Govern. Recuperat a partir de www.avaluació.cat

Casado, D., i Todeschini, F. A. (2013). Guia pràctica 10. Avaluar l'impacte de les polítiques actives d'ocupació. Ivàlua. Recuperat a partir de http://www.ivalua.cat/documents/1/17_12_2013_11_58_01_Guia10_PAO_cat.pdf

Cheung, A. C. K., i Slavin, R. E. (2012a). Effects of educational technology applications on reading outcomes for struggling readers: A best-evidence synthesis. Best Evidence Encyclopedia. Recuperat a partir de http://www.bestevidence.org/word/tech_strug_read_Jul_18_2012.pdf

Cheung, A. C. K., i Slavin, R. E. (2012b). How features of educational technology applications affect student reading outcomes: A meta-analysis. Educational Research Review, 7(3), 198-215.

Chingos, M. M., i Peterson, P. E. (2012). The Effects of School Vouchers on College Enrollment: Experimental Evidence from New York City. Brown Center on Education Policy at Brookings Institution. Recuperat a partir de http://www.hks.harvard.edu/pepg/PDF/Impacts_of_School_Vouchers_FINAL.pdf

Clark, M. A., Gleason, P. M., Tuttle, C. C., i Silverberg, M. K. (2014). Do Charter Schools Improve Student Achievement? Educational Evaluation and Policy Analysis. Online first.

Collet, J., i Tort, A. (2011). Famílies, escola i èxit: millorar els vincles per millorar els resultats. Barcelona: Fundació Jaume Bofill, Col. Informe Breus 35.

Comas, M., Escapa, S., i Abellán, C. (2014). Com participen mares i pares a l'escola? Diversitat familiar i d'implicació en educació. Barcelona: Fundació Jaume Bofill, Col. Informes Breus 49.

Cook, P. J., et al. (2014). The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago. National Bureau of Economic Research, NBER Working Paper No. 19862. Recuperat a partir de http://www.povertyactionlab.org/publication/surprising-efficacy-academic-and-behavioral-intervention-disadvantaged-youth-results-ran

Correnti, R. (2009). Examining CSR program effects on student achievement: Causal explanation through examination of implementation rates and student mobility. Ponència presentada a la 2nd Annual Conference of the Society for Research on Educational Effectiveness. Washington, DC., March 2009. Recuperat a partir de http://www.lrdc.pitt.edu/BOV/documents/Correnti_ExaminingCSRProgramEffects_033012.pdf

CREDO (2009). Multiple Choice: Charter Schools Performance in 16 States. Center for Research on Education Outcomes (CREDO), Stanford University. Recuperat a partir de http://www.btu.org/sites/default/files/research/credo_standrod_charter_school_performance_full.pdf

Currie, J., i Thomas, D. (2000). School Quality and the Longer-Term Effects of Head Start. The Journal of Human Resources, 35(4), 755-774.

Curs, B. R., i Harper, C. E. (2012). Financial aid and first-year collegiate GPA: a regression discontinuity approach. The Review of Higher Education, 35(4), 627-649.

DesJardins, S. L., McCall, B. P., Ott, M., i Kim, J. (2010). A Quasi-Experimental Investigation of How the Gates Millennium Scholars Program Is Related to College Students' Time Use and Activities. Educational Evaluation and Policy Analysis, 32(4), 456-475.

Di Carlo, M. (2011). The Evidence on Charter Schools and Test Scores. The Albert Shanker Institute Policy Brief. Recuperat a partir de http://shankerblog.cdjd.info/wp-content/uploads/2011/12/CharterReview.pdf

Dobbie, W., i Fryer Jr, R. G. (2011). Getting beneath the veil of effective schools: Evidence from New York City. National Bureau of Economic Research, NBER Working Paper No. 17632. Recuperat a partir de http://www.nber.org/papers/w17632

Dolton, P., i Marcenaro-Gutierrez, O. D. (2011). If you pay peanuts do you get monkeys? A cross-country analysis of teacher pay and pupil performance. Economic policy, 26(65), 5-55.

Duflo, E., Glennerster, R., i Kremer, M. (2007). Using randomization in development economics research: A toolkit. Handbook of development economics, 4, 3895-3962.

Duflo, E., Hanna, R., i Rya, S. P. (2012). Incentives Work: Getting Teachers to Come to School. The American Economic Review, 102(4), 1241-1278.

Duncan, G. J., i Magnuson, K. (2013). Investing in Preschool Programs. Journal of Economic Perspectives, 27(2), 109-132.

Dupriez, V., Dumay, X., i Vause, A. (2008). How Do School Systems Manage Pupils' Heterogeneity? Comparative Education Review, 52(2), 245-273.

Duru-Bellat, M., i Suchaut, B. (2005). Organisation and Context, Efficiency and Equity of Educational Systems: what PISA tells us. European Educational Research Journal, 4(3), 181–194.

Dynarski, S., i Scott-Clayton, J. (2013). Financial Aid Policy: Lessons from Research National Bureau of Economic Research, NBER Working Paper No. 18710. Recuperat a partir de http://www.nber.org/papers/w18710

Ehri, L. C., Dreyer, L. G., Flugman, B., i Gross, A. (2007). Reading Rescue: An Effective Tutoring Intervention Model for Language-Minority Students Who Are Struggling Readers in First Grade. American Educational Research Journal, 44(2), 414-448.

Fack, G., i Grenet, J. (2015). Improving College Access and Success for Low-Income Students: Evidence from a Large French Need-based Grant Program. American Economic Journal: Applied Economics, 7(2), 1-34.

Felgueroso, F., Gutiérrez-Doménech, M., i Jiménez-Martín, S. (2013). ¿Por qué el abandono escolar se ha mantenido tan elevado en España en las últimas dos décadas? El papel de la Ley de Educación (LOGSE). Economic Reports, 02-2013, 1-26.

Fryer Jr, R. G., Levitt, S. D., List, J., i Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. National Bureau of Economic Research, NBER Working Paper No. 18237. Recuperat a partir de http://www.nber.org/papers/w18237

Fryer Jr, R. G. (2011a). Creating« No Excuses»(Traditional) Public Schools: Preliminary Evidence From an Experiment in Houston. National Bureau of Economic Research, NBER Working Paper No. 17494. Recuperat a partir de http://www.nber.org/papers/w17494

Fryer Jr, R. G. (2011b). Teacher incentives and student achievement: Evidence from New York City public schools. National Bureau of Economic Research, NBER Working Paper No. 16850. Recuperat a partir de http://www.nber.org/papers/w16850

Fuchs, T., i Woessmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. Empirical Economics, 32(2-3), 433-464.

Gibbons, S., Machin, S., i Silva, O. (2008). Choice, competition, and pupil achievement. Journal of the European Economic Association, 6(4), 912–947.

Glazerman, S., Protik, A., Teh, B., Bruch, J., i Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. U.S. Department of Education, IES-NCEE 2014-4003. Recuperat a partir de http://ies.ed.gov/ncee/pubs/20144003/pdf/20144003.pdf

Glewwe, P., Ilias, N., i Kremer, M. (2010). Teacher incentives. American Economic Journal: Applied Economics, 2(3), 205-227.

Goos, M., Schreier, B. M., Knipprath, H. M. E., De Fraine, B., Van Damme, J., i Trautwein, U. (2013). How Can Cross-Country Differences in the Practice of Grade Retention Be Explained? A Closer Look at National Educational Policy Factors. Comparative Education Review, 57(1), 54-84.

Gorard, S., See, B. H., i Siddiqui, N. (2014). Switch-on Reading. Education Endowment Foundation. Recuperat a partir de http://educationendowmentfoundation.org.uk/uploads/pdf/FINAL_EEF_Evaluation_Report_-_Switch-on_-_February_2014.pdf

Gorard, S., i Smith, E. (2004). An international comparison of equity in education systems. Comparative education, 40(1), 15-28.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short-and long-term benefits of educational opportunity. School Psychology Quarterly, 16(1), 9.

Goux, D., Gurgand, M., i Maurin, E. (2013). The effect of school and peers on droput behavior. Ivàlua. Recuperat a partir de http://www.ivalua.cat/documents/1/28_06_2013_07_44_49_Marc_Gurgand.pdf

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. Educational Evaluation and Policy Analysis, 21(2), 143-163.

Hanushek, E. A., Link, S., i Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. Journal of Development Economics, 104, 212–232.

Hanushek, E. A., i Rivkin, S. G. (2007). Pay, working conditions, and teacher quality. The future of children, 17(1), 69–86.

Hanushek, E. A., i Woessmann, L. (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. The Economic Journal, 116(510), C63–C76.

Hanushek, E. A., i Woessmann, L. (2010). The economics of international differences in educational achievement. National Bureau of Economic Research, NBER Working Paper No. 15949. Recuperat a partir de http://www.nber.org/papers/w15949

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. International Journal of Educational Research, 43(6), 387-425.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., i Yavitz, A. (2010a). The rate of return to the HighScope Perry Preschool Program. Journal of Public Economics, 94(1–2), 114-128.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P., i Yavitz, A. (2010b). A new cost-benefit and rate of return analysis for the Perry Preschool Program: A summary. National Bureau of Economic Research, NBER Working Paper No. 16180. Recuperat a http://www.nber.org/papers/w16180

Heckman, J. J., Pinto, R., i Savelyev, P. A. (2012). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. National Bureau of Economic Research, NBER Working Paper No. 18581. Recuperat a partir de http://www.nber.org/papers/w18581

Heppen, J. B., Walters, K., Clements, M., Faria, A.-M., Tobey, C., Sorensen, N., i Culp, K. (2012). Access to Algebra I: The Effects of Online Mathematics for Grade 8 Students. U.S. Department of Education, IES-NCEE 2012-4021. Recuperat a partir de http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_20124021.pdf

Hew, K. F., i Cheung, W. S. (2013). Use of Web 2.0 technologies in K-12 and higher education: The search for evidence-based practice. Educational Research Review, 9, 47-64.

Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. A L. A. Shepard i M. L. Smith (Eds.), Flunking grades: Research and policies on retention (pp. 16-33). London: The Falmer Press.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. The Quarterly Journal of Economics, 115(4), 1239-1285.

Hoxby, C. M., Murarka, S., i Kang, J. (2009). How New York City's charter schools affect achievement. Cambridge, MA: The New York City Charter Schools Evaluation Project 2009. Recuperat a partir de http://users.nber.org/~schools/charterschoolseval/how_NYC_charter_schools_affect_achievement_sept2009.pdf

Imbens, G. W. (2014). Instrumental Variables: An Econometrician's Perspective. National Bureau of Economic Research, NBER Working Paper No. 19983. Recuperat a partir de http://www.nber.org/papers/w19983

Jacob, B. A., i Lefgren, L. (2009). The effect of grade retention on high school completion. American Economic Journal: Applied Economics, 1(3), 33–58.

Jacob, R. T., Smith, T. J., Willard, J. A., i Rifkin, R. E. (2014). Reading Partners: The Implementation and Effectiveness of a One-on-One Tutoring Program Delivered by Community Volunteers. MDRC Policy Brief. Recuperat a partir de http://readingpartners.org/wp-content/uploads/2014/06/Reading-Partners_final.pdf

Jeynes, W. H. (2005). A Meta-Analysis of the Relation of Parental Involvement to Urban Elementary School Student Academic Achievement. Urban Education, 40(3), 237-269.

Jeynes, W. H. (2007). The Relationship Between Parental Involvement and Urban Secondary School Student Academic Achievement A Meta-Analysis. Urban Education, 42(1), 82-110.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. School Psychology Review, 30(3), 420–437.

Kidron, Y., i Lindsay, J. (2014). The Effects of Increased Learning Time on Student Academic and Nonacademic Outcomes: Findings from a Meta-Analytic Review. REL 2014-015. IES-NCEER / REL Appalachia. Recuperat a partir de http://eric.ed.gov/?id=ED545233

Lavy, V. (2012). Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behavior. National Bureau of Economic Research, NBER Working Paper No. 18369. Recuperat a partir de http://www.nber.org/papers/w18369

Leeds, D. M., i DesJardins, S. L. (2014). The Effect of Merit Aid on Enrollment: A Regression Discontinuity Analysis of Iowa's National Scholar's Award. Mimeo. Recuperat a partir de https://aefpweb.org/sites/default/files/webform/aefp40/The%20Effect%20of%20Merit%20Aid%20on%20Enrollment.pdf

Lee, Y. S., Morrow-Howell, N., Jonson-Reid, M., i McCrary, S. (2012). The Effect of the Experience Corps® Program on Student Reading Outcomes. Education and Urban Society, 44(1), 97-118.

Loeb, S., i Lee, V. E. (1995). Where do Head Start attendees end up? One reason why preschool effects fade out. Evaluation and Policy Analysis, 17(1), 62-82.

Loeb, S., i Page, M. E. (2000). Examining the link between teacher wages and student outcomes: The importance of alternative labor market opportunities and non-pecuniary variation. Review of Economics and Statistics, 82(3), 393-408.

Manacorda, M. (2008). The cost of grade retention. CEP Discussion Paper No. 878. London: Centre for Economic Performance, London School of Economics and Political Science. Recuperat a partir de http://cep.lse.ac.uk/pubs/download/dp0878.pdf

Manacorda, M. (2012). The cost of grade retention. Review of Economics and Statistics, 94(2), 596–606.

Manning, M., Homel, R., i Smith, C. (2010). A meta-analysis of the effects of early developmental prevention programs in at-risk populations on non-health outcomes in adolescence. Children and Youth Services Review, 32(4), 506-519.

Mattingly, D. J., Prislin, R., McKenzie, T. L., Rodriguez, J. L., i Kayzar, B. (2002). Evaluating evaluations: The case of parent involvement programs. Review of Educational Research, 72(4), 549-576.

Mayer, D. P., Peterson, P. E., Myers, D. E., Tuttle, C. C., i Howell, W. G. (2002). School choice in New York City after three years: An evaluation of the school choice scholarships program. Mathematica Policy Research, MPR No. 8404-045. Recuperat a partir de http://www.mathematica-mpr.com/~/media/publications/PDFs/nycfull

May, H., et al. (2014). Evaluation of the i3 Scale-up of Reading Recovery. CPRE, No. RR-79. Recuperat a partir de http://lfws.literacy.org/sites/default/files/researchreport/1488_readingrecoveryreport.pdf

McGiverin, J., Gilman, D., i Tillitski, C. (1989). A meta-analysis of the relation between class size and achievement. The Elementary School Journal, 90(1), 47-56.

Mediavilla, M. (2012). Les beques com a factor d'èxit escolar: Una avaluació d'impacte a partir d'una metodologia quasi-experimental. A M.Martínez i B. Albaigés (Eds.) L'estat de l'educació a Catalunya. Anuari 2011 (pp. 167-182). Barcelona: Mediterrània, Col. Polítiques 75.

Mediavilla, M. (2013). Heterogeneidad en el impacto de la política de becas en la escolaridad secundaria postobligatoria en España: un análisis por subgrupos poblacionales. Estudios de Economia, 40(1), 97-120.

Meghir, C., i Palme, M. (2005). Educational reform, ability, and family background. The American Economic Review, 95(1), 414-424.

Miller, S., i Connolly, P. (2013). A Randomized Controlled Trial Evaluation of Time to Read, a Volunteer Tutoring Program for 8- to 9-Year-Olds. Educational Evaluation and Policy Analysis, 35(1), 23-37.

Miller, S., Connolly, P., i Maguire, L. K. (2012). The effects of a volunteer mentoring programme on reading outcomes among eight- to nine-year-old children: A follow up randomized controlled trial. Journal of Early Childhood Research, 10(2), 134-144.

Morris, S., Tödling-Schönhofer, H., i Wiseman, M. (2013). Design and commissioning of counterfactual impact evaluations guidance to help employers and workers to manage the transition to the new classification, labelling and packaging system. Luxembourg: European Commission, DG Employment, Social Affairs and Inclusion.

Nelson, G., Westhues, A., i MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. Prevention & Treatment, 6(1), 31a.

OECD. (2007). PISA 2006 Science Competencies for Tomorrow's World. Vol. 1. Paris: OECD Publishing. Recuperat a partir de http://www.navarro.cl/web/educacion/docs_comision/Documentos%20e%20informes%20academicos/PISA.pdf

OECD. (2010). PISA 2009 Results: What Students Know and Can Do. Vol. 4. Paris: OECD Publishing. Recuperat a partir de http://www.llv.li/rss/pdf-llv-sa-pisa_2009_oecd-bericht_englisch__band_1_-_was_schueler_wissen_und_koennen.pdf

OECD. (2014). PISA 2012 Results. What Makes Schools Successful? Vol. 1. How Resources, Policies and Practices are Related to Education Outcomes. Paris: OECD Publishing. Recuperat a partir de http://www.oecd.org/pisa/keyfindings/Vol4Ch1.pdf

Patall, E. A., Cooper, H., i Allen, A. B. (2010). Extending the School Day or School Year: A Systematic Review of Research (1985-2009). Review of Educational Research, 80(3), 401-436.

Peck, L. R., i Bell, S. H. (2014). The Role of Program Quality in Determining Head Start's Impact on Child Development. Third Grade Follow-Up to the Head Stard Impact Study. OPRE Report No. 2014-10. Recuperat a partir de http://www.acf.hhs.gov/sites/default/files/opre/hs_quality_report_4_28_14_final.pdf

Pekkarinen, T., Uusitalo, R., i Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. Journal of Public Economics, 93(7), 965-973.

Porter, S. R. (2012). Using Instrumental Variables Properly to Account for Selection Effects. Mimeo. Recuperat a partir de http://eric.ed.gov/?id=ED531905

Puma, M., et al. (2012). Third Grade Follow-Up to the Head Start Impact Study: Final Report. OPRE Report No. 2012-45. Recuperat a partir de http://eric.ed.gov/?id=ED539264

Redd, Z., Boccanfuso, C., Walker, K., Princiotta, D., Knewstub, D., i Moore, K. (2012). Expanding time for learning both inside and outside the classroom: A review of the evidence base. Child Trends. Recuperat a partir de http://childtrends.org/wp-content/uploads/2013/03/Child_Trends-2012_08_16_RB_TimeForLearning.pdf

Rubin, R. B. (2011). The Pell and the Poor: A Regression-Discontinuity Analysis of On-Time College Enrollment. Research in Higher Education, 52(7), 675-692.

Rutt, S., Easton, C., i Stacey, O. (2014). Catch Up® Numeracy. Education Endowment Foundation. Recuperat a partir de http://www.nfer.ac.uk/publications/EFCU01/EFCU01.pdf

Schlotter, M., Schwerdt, G., i Woessmann, L. (2010). Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide. IZA DP No. 4725. Recuperat a partir de http://www.eenee.de/portal/page/portal/EENEEContent/_IMPORT_TELECENTRUM/DOCS/EENEE_AR5.pdf

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., i Shavelson, R. J. (2007). Estimating causal effects: Using experimental and observational designs. American Educational Research Association, Washington, DC. Recuperat a partir de http://69.8.231.237/uploadedFiles/Publications/Books/Estimating_Causal_Effects/ECE_Front-TOC.pdf

Schütz, G. (2009). Does the quality of pre-primary education pay off in secondary school? An international comparison using PISA 2003. Ifo Working Paper 68. Recuperat a partir de http://www.cesifo-economic-studies.de/pls/guest/download/Ifo%20Working%20Papers%20%28seit%202005%29/IfoWorkingPaper-68.pdf

Schütz, G., Ursprung, H. W., i Woessmann, L. (2008). Education policy and equality of opportunity. Kyklos, 61(2), 279-308.

Schütz, G., West, M., i Woessmann, L. (2007). School accountability, autonomy, choice, and the equity of student achievement: International evidence from PISA 2003. Paris: OECD Publishing. Recuperat a partir de http://www.oecd.org/edu/39839422.pdf

Scott-Clayton, J. (2011). On Money and Motivation A Quasi-Experimental Analysis of Financial Incentives for College Achievement. Journal of Human Resources, 46(3), 614-646.

Slavin, R. E. (1989). Class size and student achievement: Small effects of small classes. Educational Psychologist, 24(1), 99-110.

Slavin, R. E., Lake, C., Davis, S., i Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. Educational Research Review, 6(1), 1-26.

Springer, M. G., et al. (2010). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). RAND Educaction i National Center of Performance Incentives. Recuperat a partir de http://www.rand.org/content/dam/rand/pubs/reprints/2010/RAND_RP1416.pdf

Sundararaman, V., i Muralidharan, K. (2011). Teacher Performance Pay: Experimental Evidence from India. The Journal of Political Economy, 119(1), 39-77.

Thernstrom, A., i Thernstrom, S. (2004). No excuses: Closing the racial gap in learning. New York: Simon and Schuster.

Van Steensel, R., McElvany, N., Kurvers, J., i Herppich, S. (2011). How effective are family literacy programs? Results of a meta-analysis. Review of Educational Research, 81(1), 69-96.

Van Voorhis, F. L., Maier, M. F., Epstein, J. L., Lloyd, C., i Leung, T. (2013). The impact of family involvement on the education of children age 3 to 8. MDRC Report. Recuperat a partir de http://www.mdrc.org/sites/default/files/The_Impact_of_Family_Imvolvement_ES.pdf

Whitman, D. (2008). Sweating the small stuff: Inner-city schools and the new paternalism. Washington, DC.: Thomas B. Fordham Institute

Woessmann, L. (2003). Schooling resources, educational institutions and student performance: the international evidence. Oxford Bulletin of Economics and Statistics, 65(2), 117-170.

Woessmann, L. (2005a). Educational production in Europe. Economic policy, 20(43), 445-504.

Woessmann, L. (2005b). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. Education Economics, 13(2), 143-169.

Woessmann, L., Luedemann, E., i Schuetz, G. (2009). School accountability, autonomy and choice around the world. Cheltenham: Edward Elgar Publishing Ltd.

Woessmann, L., Luedemann, E., Schuetz, G., i West, M., R. (2007). School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003. Paris: OECD Publishing. Recuperat a partir de http://www.oecd.org/edu/39839361.pdf

Woessmann, L., i West, M. R. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. European Economic Review, 50(3), 695-736.

York, B. N., i Loeb, S. (2014). One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents of Preschoolers. National Bureau of Economic Research. NBER Working Paper No. 20659. Recuperat a partir de http://www.nber.org/papers/w20659

Zimmer, R. W., Gill, B., Booker, K., Lavertu, S., Sass, T. R., i Witte, J. (2009). Charter schools in eight states: Effects on achievement, attainment, integration and competition. Santa Monica: RAND Corporation.

ivàlua
Institut Català d'Avaluació de Polítiques Públiques
Pau Claris, 108
4t 1a
08009 Barcelona
T. +34 93 554 53 00
info@ivalua.cat
www.ivalua.cat

Ejemplar impreso: 12€
Versión electrónica: www.ivalua.cat