



Facultad
de Ciencias
Económicas y
Empresariales

Departamento
de Economía
Aplicada y
Estadística



**Revista de Evaluación de
Programas y Políticas Públicas**
JOURNAL OF PUBLIC PROGRAMS AND POLICY EVALUATION

La perspectiva «qué funciona» en educación: algunas aplicaciones

Núm. 2 (2014), pp. 155-186.

Miquel Àngel Alegre Canosa¹

Recibido: Enero, **2014**

Aceptado: Febrero, **2014**

JEL Clasif: JEL-class: I21, I24, I28

¹ Este artículo toma en consideración parte del contenido del informe (inédito) “Què funciona en educació: el repte d’avaluar l’efectivitat de les polítiques educatives”, cofinanciado por la Obra Social “La Caixa” en el marco del convenio entre la Obra Social de “La Caixa” e Ivàlua. Ivàlua e IGOP. Correspondencia.

Miquel Àngel Alegre, IGOP y Ivàlua, miquelangel.alegre@ivalua.cat

Abstract

This article echoes the perspective of “what works” in education and shows how it could be applied in the design and implementation of educational policies. Firstly, we outline the origins of ‘what works’ in education, with an emphasis on the school effectiveness research tradition. Secondly, we discuss what kind of evidence is needed to show that an educational intervention is effective. Finally, on the basis of this evidence, we take the case of educational inequalities by selecting some illustrative examples of educational policies pursuing this aim.

Key Words: education, evaluation, effectiveness, impact, evidence, what works.

Resumen

Este artículo discute y desarrolla la aplicación de la perspectiva «qué funciona» en el ámbito de la evaluación de políticas y programas educativos. Este objetivo se desgana en tres ejercicios. En primer lugar, se presentan los antecedentes más inmediatos de esta perspectiva, en particular la tradición de investigación sobre efectividad escolar. En segundo lugar, se identifica el tipo de evidencia que la perspectiva «qué funciona» establece como prueba de la efectividad de las intervenciones educativas. En tercer lugar, y partiendo de este mismo criterio, se revisa la evidencia acumulada en torno a la capacidad de una selección (no exhaustiva) de actuaciones educativas de reducir las desigualdades educativas.

Palabras clave: educación, evaluación, efectividad, impacto, evidencia, qué funciona.

1. Introducción

De forma general, la perspectiva «qué funciona» (what works) de análisis y evaluación de políticas y programas tiene su razón de ser en la apuesta por avanzar hacia decisiones y políticas basadas en la evidencia (evidence-based policy making). Este artículo aborda esta perspectiva en el ámbito de las políticas educativas, y lo hace a través de un triple ejercicio. En primer lugar, se rastrean los antecedentes más inmediatos del enfoque «qué funciona» en la investigación educativa, y se presentan algunos de sus avances en el terreno metodológico e institucional. En segundo lugar, y dado que la perspectiva «qué funciona» está dirigida principalmente a proporcionar evidencias de impactos, discutiremos sobre el concepto de evidencia en sí mismo en el ámbito de la educación. Más específicamente, nos preguntaremos qué es evidencia sólida y creíble y qué no lo es, y caracterizaremos muy brevemente el tipo de prueba empírica que la perspectiva en cuestión reclama cuando se trata de establecer relaciones causales entre intervención educativa y cambios en la problemática a tratar. En tercer lugar, pasaremos breve revista a la práctica evaluativa y a la evidencia acumulada en torno a una selección de ámbitos estratégicos de la política y la intervención educativa, pasando del nivel más contextual (hablaremos aquí de las políticas de autonomía escolar) al más micro (programas y prácticas educativas en las escuelas). Es necesario advertir que esta selección no tiene ningún ánimo de exhaustividad, sino que se realiza teniendo en cuenta, primero, la posibilidad de que la muestra escogida acabe ofreciendo un conjunto ilustrativo de ejemplos especialmente relevantes de enfoques de evaluación y, segundo, las prioridades y los puntos calientes que forman parte de la agenda de la política educativa en nuestro país. Además, esta revisión se centra en aquellas evidencias que nos hablan de la capacidad de las políticas y los programas para incidir en un objetivo concreto: la reducción de las desigualdades sociales en educación. Nos fijamos, por tanto, en la efectividad de unas intervenciones y otras a la hora de ampliar las oportunidades educativas de los alumnos socialmente más desfavorecidos.

2. Origen y alcance de la perspectiva «qué funciona» en educación

En el ámbito educativo, podemos considerar el enfoque «qué funciona» como la última evolución de la denominada *school effectiveness research* (SER) o investigación sobre la efectividad escolar. Esta línea de investigación nació a principios de los años ochenta como reacción a la fuerza de una literatura que, amparándose en las conclusiones del conocido informe (Coleman, 1968), ponía el énfasis en la composición socioeconómica y étnica de las escuelas como el principal factor con capacidad para condicionar el rendimiento académico de los estudiantes (más allá de los recursos de que disponen y de lo que hacen con ellos). Los estudios de la SER, por el contrario, tratan de demostrar que los centros educativos, la manera en que trabajan y se organizan, pueden contribuir a mejorar los resultados de sus alumnos, llegando incluso a reducir las desigualdades académicas entre unos y otros (Teddlie & Reynolds, 2000). Dicho de otra forma, que las escuelas, al margen de su composición escolar, «can make the difference». Algunas de las características de centro que esta literatura ha identificado como particularmente efectivas son, por ejemplo, una atmósfera de consenso, cooperación y cohesión entre el profesorado, la existencia de un clima escolar ordenado o la orientación de la práctica docente hacia el éxito académico de todos los estudiantes (Scheerens & Bosker, 1997). Desde sus inicios, estos estudios han ido refinando sus planteamientos metodológicos, por ejemplo, incorporando en sus análisis variables relacionadas con la composición social de las escuelas, una opción que la SER había menospreciado tradicionalmente (Thrupp, 2001).

La perspectiva «qué funciona» en educación comparte con la SER el objetivo de estudiar la efectividad de las intervenciones educativas, así como el interés por aportar evidencia empírica a los procesos de toma de decisiones en política educativa. No obstante, ambas perspectivas se diferencian en un elemento central: el tipo de evidencia que unas y otras tienden o aspiran a generar. Por un lado, la tradición de la SER se compone principalmente de estudios correlacionales. Es decir, utilizando datos observacionales y mediante diversas técnicas estadísticas multivariantes, proporcionan evidencia sobre el grado de asociación existente entre variables independientes (intervenciones o determinadas características escolares) y variables dependientes (por ejemplo, rendimiento académico de alumnos y escuelas) (Luyten, H., Visscher, A., &

Witziers, B., 2005). La perspectiva «qué funciona», en cambio, aspira a evidenciar relaciones causales entre la intervención y los outcomes considerados, lo cual ha supuesto, en el terreno de la evaluación educativa, una apuesta clara por diseños robustos de evaluación de impactos, principalmente diseños experimentales.

Sin embargo, hace muy poco tiempo que el campo de las políticas educativas ha empezado a abrir la puerta a estos últimos avances en el terreno de la evaluación, y ha ocurrido, principalmente, en el mundo anglosajón. Mencionamos aquí dos iniciativas especialmente relevantes, ejemplos de institucionalización de la perspectiva «qué funciona» en el ámbito de la política educativa. El primero lo situamos en los Estados Unidos, en la iniciativa What Works Clearinghouse (WWC), impulsada por el Institute of Education Sciences del Gobierno federal de los Estados Unidos. La iniciativa nació en el año 2002, vinculada al programa No Child Left Behind y al énfasis que pone en la necesidad de que las actuaciones educativas sufragadas con financiación pública dispongan de —y también generen— pruebas de efectividad. Así pues, el WWC revisa la evidencia existente sobre la capacidad de impacto de los programas educativos, valora la calidad de esta evidencia de acuerdo con estándares muy exigentes y, sobre esta base, establece el grado de efectividad de los programas en cuestión. Toda esta información se pone al alcance de responsables políticos, investigadores y miembros de la comunidad educativa, con el objetivo de promover decisiones y prácticas basadas en la evidencia.

El segundo ejemplo lo encontramos en el Reino Unido, en la constitución de la Education Endowment Foundation (EEF) en el año 2011. Esta fundación, reconocida por el gobierno británico como el centro what works especializado en el ámbito educativo, concentra su misión en la generación de evidencias en torno a lo que funciona (y no funciona) en la lucha contra las desigualdades educativas. Una de las principales herramientas que desarrolla la EEF es el denominado Teaching and Learning Toolkit, donde se resume y califica 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. Así pues, la EEF, y también la WWC, desempeñan una función de interfaz o mediación (brokerage) entre el mundo académico y el mundo de la decisión política, facilitando la adaptación del tiempo y del lenguaje típico del primero a las necesidades de inmediatez y claridad del segundo.

En España, como en tantos otros países de nuestro entorno, el campo de la política educativa está aún lejos de lo que suponen estos avances. Como veremos, las decisiones sobre las políticas educativas —sobre su lanzamiento, su mantenimiento, su reforma, su supresión— raramente se basan en evidencias empíricas sólidas sobre su posible efectividad; de hecho, no es infrecuente que tras estas decisiones no haya evidencia empírica de ningún tipo. Contribuciones como la que aquí se presenta pretenden ayudar a contrapesar esta tendencia.

3. ¿De qué evidencia estamos hablando? Evaluación experimental y cuasi-experimental

Así pues, la perspectiva «qué funciona» pone su atención en la efectividad de las políticas, es decir, en los impactos que estas políticas son capaces de provocar sobre las problemáticas que abordan. No obstante, la pregunta «qué funciona» no admite cualquier respuesta, sino únicamente aquellas que se fundamentan en evidencias empíricas creíbles y sólidas, las que establecen de forma robusta las relaciones de causalidad entre la política y eventuales cambios en la problemática abordada. Llegados a este punto, habría que preguntarse: ¿qué deberíamos considerar evidencia empírica creíble y sólida en el campo educativo? ¿Qué tipo de evidencias necesitamos obtener o acumular para determinar con garantías si una política o programa educativo funciona o no funciona, si es más o menos efectivo en la lucha contra las desigualdades educativas?

Obviamente, este artículo no puede aportar una respuesta mínimamente exhaustiva a estos interrogantes. Otros títulos recientes ofrecen panorámicas y ejemplos interesantes de diseños y métodos de evaluación capaces de proporcionar evidencias robustas de la efectividad (o inefectividad) de las políticas y los programas educativos (Schlotter, Schwerdt, & Woessmann, 2010; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Nos conformamos aquí con las notas que vienen a continuación, primero sobre la evaluación experimental, segundo sobre uno de los métodos cuasiexperimentales de mayor proyección en el ámbito que nos ocupa.

La forma más robusta de aproximarse a los impactos de una intervención educativa pasa por aleatorizar quién acaba siendo objeto de ella, es decir, por la aplicación del

mecanismo de sorteo como procedimiento para asignar, de entre el conjunto de unidades elegibles para el tratamiento (alumnos, clases, escuelas), cuáles acaban recibiendo y cuáles se incorporan al grupo de control. Este procedimiento es la piedra angular de la evaluación experimental de impacto, un diseño evaluativo 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 el número de casos es lo suficientemente elevado, la aleatorización garantiza que el conjunto de unidades control y el conjunto de tratamiento acaban siendo equivalentes en todas aquellas características, observables y no observables, que pueden estar relacionadas con la consecución del *outcome* de interés. Lo único que les diferenciará entonces será el hecho de haber participado o no en el programa, causa de toda posible diferencia en los resultados correspondientes. Más particularmente, el que la asignación aleatoria se realice entre solicitantes activos de un determinado recurso o servicio permite comparar conjuntos de individuos (o escuelas, según el caso) con un mismo nivel de motivación (fuente de inobservables por *antonomasia*), al menos el nivel de motivación vinculado al acto de solicitarlo.

El hecho de que la definición de partida de las características que deben reunir los individuos o escuelas solicitantes de la ayuda (población elegible) sea más o menos restringida no tiene por qué cuestionar la robustez de los resultados de la evaluación, su validez interna. Más aún, cuando se restringe el perfil de individuos o escuelas elegibles para la intervención (población especialmente desfavorecida, por ejemplo) y cuando su número supera la capacidad de cobertura del programa, la aleatorización de la participación es, seguramente, el mecanismo de distribución más justo que podamos diseñar. Aparece, eso sí, un cierto *trade-off* entre el nivel de focalización de la intervención y la validez externa de los resultados de la evaluación (es decir, en qué medida estos pueden ser extrapolables a contextos y poblaciones diferentes). En cualquier caso, es evidente que las evaluaciones experimentales plantean retos importantes tanto desde un punto de vista metodológico como en términos de viabilidad política. Sin embargo, como ya se ha expuesto en otros lugares (Casado, 2012), lo sorprendente es que este diseño evaluativo —utilizado profusamente en el extranjero y considerado el método más robusto de evaluación de impacto— no se haya utilizado aún en nuestro país para evaluar ninguna política pública, tampoco las educativas.

Cuando la aplicación de diseños experimentales, por unos u otros motivos, no es viable, pueden entrar en juego determinados diseños de evaluación cuasiexperimental. Estos métodos comparten con el diseño experimental la definición de un grupo de control (no participantes en el programa) también utilizado como escenario 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.

Variabes instrumentales, regresiones discontinuas, dobles diferencias, técnicas de panel y uso de efectos fijos, propensity score matching..., no son pocos los métodos no experimentales que han sido y continúan siendo utilizados para evaluar la efectividad de las intervenciones educativas. Uno de ellos, el diseño de regresión discontinua, ha experimentado una proyección más que notable en los últimos años. De hecho, esta técnica gana vigor en el marco de lo que se ha denominado experimentos naturales, plausibles cuando el hecho de encontrarse expuesto o no a una determinada política depende de una circunstancia exógena (cuasialeatoria desde el punto de vista de participantes y controles), no vinculada al outcome de interés.

La regresión discontinua se aplica cuando la participación en el programa a evaluar 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. Así, por ejemplo, podría darse el caso de que un plan de priorización educativa otorgase una dotación extraordinaria de recursos a todos los centros que tuvieran más de un 25 % de alumnos con déficits instructivos graves. En tanto que la definición de estos umbrales ha sido establecida de forma exógena (en este caso, diríamos, la administración educativa), cabría esperar que los centros y los alumnos que se encuentren justo por debajo y justo por encima del punto de corte sean esencialmente equivalentes en todos aquellos atributos que puedan resultar relevantes, tanto observables como inobservables. Si el umbral se aplica de manera estricta, lo que se verá es un salto en el porcentaje de participantes justo por encima y justo por debajo del punto de corte en cuestión. Esto es, debería observarse que el porcentaje de centros con una dotación extraordinaria de recursos presenta un salto cuando estos pasan a tener más de un 25 % de alumnos con déficits académicos. De esta forma, si el programa es efectivo, deberíamos poder observar que el salto en el porcentaje de

participación se corresponde con otra discontinuidad entre el nivel de rendimiento de los centros que se encuentran justo por debajo del 25 % y el de los que se encuentran justo por encima. 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. A lo largo de este artículo veremos ejemplos de evaluaciones de programas educativos basadas en el uso de esta técnica.

4. Qué funciona y cómo lo sabemos: algunos ejemplos de interés

Presentamos aquí una revisión de la práctica evaluativa y de la evidencia que la investigación internacional ha ido acumulando en torno a una selección de ámbitos clave de la política educativa. Recordamos que esta selección no es ni pretende ser exhaustiva, y que concentra su atención en la capacidad de las políticas educativas de mejorar las perspectivas educativas del alumnado socialmente más desfavorecido.

4.1. Un elemento de contexto: el debate sobre la autonomía escolar

Acompañando el impulso de las políticas en favor de la elección escolar, la delegación de más autonomía a los centros escolares ha pasado a ocupar una posición central en el debate educativo global. Entre otras posibles ventajas, se concibe que una mayor autonomía pedagógica y de funcionamiento debería permitir a los centros ajustarse mejor a las particularidades de su población escolar y, por tanto, afinar la atención a sus diversidades. En Cataluña, por ejemplo, habría que recordar el lanzamiento en 2005 del Proyecto de Autonomía de Centros (PAC), prueba piloto que incorporaba la descentralización a las escuelas de una serie de funciones relacionadas con los ámbitos pedagógico, organizativo y de gestión de recursos humanos y materiales, introduciendo también los correspondientes mecanismos de seguimiento y de rendición de cuentas. Años después, la ley de Educación de Cataluña (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 (2010) de trasladar los criterios del PAC al conjunto de centros públicos y concertados de Cataluña. Tampoco en este caso podemos afirmar que la introducción de reformas haya

venido motivada por la disposición de evidencias empíricas. En particular, cabría interpretar la falta de una evaluación de impacto del PAC —en este caso especialmente, ya que el plan se planteaba en términos de prueba piloto— como una oportunidad perdida de someter a prueba la efectividad de unas medidas tan importantes como estas antes de ser generalizadas.

¿Qué nos dice la evidencia internacional sobre la efectividad de los programas de autonomía escolar? Por una parte, un buen número de estudios basados en pruebas internacionales de competencias (principalmente PISA, TIMSS o PIRLS²) 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 gradiente socioeconómico del rendimiento se incrementa en los países donde una parte importante de las escuelas pueden interceder en el proceso de selección y admisión de los alumnos y, por tanto, caer en la tentación del *cream-skimming*³ (Alegre & Ferrer, 2010). Finalmente, algunos de estos estudios evidencian una clara interacción entre la efectividad de la autonomía escolar y determinados mecanismos de rendición de cuentas (*accountability*), apoyando el argumento de que la autonomía es positiva —en términos de nivel medio y de equidad de resultados— siempre y cuando las escuelas estén sujetas a dichos mecanismos (Woessmann, 2005b).

Abrimos aquí un breve paréntesis. En España, la introducción de pruebas externas de competencias en momentos clave de la educación primaria y secundaria —algunas promovidas por el Gobierno central, otras por los Gobiernos autonómicos—, a pesar de tener un carácter fundamentalmente informativo, no siempre ha sido bien recibida por determinados sectores de la comunidad educativa. Entre otras cosas, se cuestionan las consecuencias perversas que estos mecanismos pueden acabar teniendo en la práctica

² PISA = Programme for International Student Assessment; TIMSS = Trends in International Mathematics and Science Study; PIRLS = Progress in *International Reading Literacy Study*.

³ Prácticas selectivas dirigidas a captar un determinado perfil de alumnado, típicamente el de más alto perfil académico y/o socioeconómico.

docente, fomentando estrategias de enseñanza principalmente orientadas a la obtención de buenos resultados en estas pruebas; es el denominado *teaching to test*. Este argumento ha cobrado fuerza como fuente de crítica al examen de reválida previsto por la LOMCE, vinculante a los efectos de obtención de los títulos de ESO y bachillerato. Algunos autores remarcan que los riesgos del *teaching to test* y de la posibilidad de que esta práctica desatienda objetivos educativos importantes, dependerán de cuál sea el diseño de contenidos de las pruebas en cuestión.⁴ De forma general, cuanto más prioricen estas pruebas la aproximación por competencias (y no por contenidos estancos), menos margen dejarán a la posibilidad de que determinadas mejoras en unas áreas curriculares se produzcan a expensas de progresos en otros conocimientos relevantes (Rychen & Salganik, 2003).

Dicho esto, recordamos que hay que leer con cautela los resultados y las conclusiones a los que llegan los estudios correlacionales. Por una parte, convendría no olvidar que la información en que se basan procede mayoritariamente de las respuestas de los directores de los centros seleccionados a unos determinados cuestionarios externos y estandarizados. Sin duda, la naturaleza subjetiva de la información, unida al instrumento utilizado para obtenerla, puede ser fuente de sesgos importantes. Podría suceder, por ejemplo, que los directores de las escuelas más exitosas presenten una mayor tendencia a autoatribuirse más capacidad de actuación autónoma que los directores de aquellas otras que, a pesar de tener el mismo reconocimiento institucional que las primeras, no consigan ser tan efectivas. Por otra parte, recordamos que la evidencia empírica que aporta esta literatura comparada es eminentemente descriptiva y que, por tanto, no identifica necesariamente relaciones de causalidad.

Una manera de eludir estas limitaciones pasa por aprovechar la variabilidad que este factor puede presentar dentro de algunos países o regiones. Efectivamente, no resulta extraño encontrar, dentro del mismo sistema educativo, tipos de centros con regímenes institucionales diferentes y con niveles de autonomía real (no únicamente percibida) también diversos. Al mismo tiempo, puede aprovecharse el hecho, tampoco infrecuente, de que algunos de estos centros —en especial, aquellos con mayor nivel de autonomía— dispongan de sistemas de admisión específicos, incorporando a menudo un mecanismo de sorteo. Un caso paradigmático de esta situación es el de las charter

⁴ Véase la entrada de Antonio Cabrales en *Nada es Gratis*, <http://www.fedeablogs.net/economia/?p=33475>.

schools, creadas en los Estados Unidos a principios de la década de 1990. El mecanismo de acceso a estas escuelas brinda una muy buena oportunidad para el diseño de experimentos dirigidos a evaluar la posible efectividad de la autonomía escolar (entre otras variables escolares).

Si bien las charter schools son centros de iniciativa y gestión fundamentalmente privadas (tercer sector o compañías empresariales), se les define como parte integrante del sistema de la educación pública. Buena parte de su financiación procede de fondos públicos, a cambio de que garanticen la gratuidad de la matriculación, y no pueden lucrarse en ningún caso. Hasta aquí, se observan claras similitudes entre las *charter* y nuestros colegios concertados. Interesa destacar dos particularidades específicas del modelo *charter*. En primer lugar, estos centros —que pueden ser de primaria o secundaria— disponen de un nivel de autonomía escolar (profesorado, presupuesto, currículum y pedagogía) muy superior al de las escuelas públicas convencionales. En segundo lugar, a pesar de que acostumbran a estar situados en zonas urbanas especialmente desfavorecidas, su elección no está sujeta a ninguna restricción de asignación zonal; en caso de exceso de demanda, el acceso viene determinado por un sorteo (lotería). Esta circunstancia permite la identificación de un grupo de control (los «perdedores» de la lotería) 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 escuela especial como lo son las *charter schools*. Así pues, no resulta extraño que estos centros hayan sido objeto de numerosas evaluaciones de efectividad; no solamente lo promueve el contrato de rendición de cuentas que estos centros firman con el estado, sino que también lo facilita el diseño del mecanismo para acceder a ellos. ¿Qué dicen los estudios sobre la efectividad de este modelo de escuelas autónomas?

Revisiones de la evidencia existente, así como evaluaciones de efectividad de las *charter schools* en más de una localidad o un estado, señalan que los impactos de este tipo de escuela suelen ser muy variables. Se constata que las *charter* funcionan en determinados contextos, pero en otros no, funcionan para determinados grupos de alumnos, pero no para otros (Di Carlo, 2011; Betts & Tang, 2011). Tanto en los casos en que los impactos sobre los *outcomes* de interés (típicamente, rendimiento académico y niveles de graduación) son positivos como en los casos en que son negativos, siempre son ganancias y pérdidas (respecto de los resultados obtenidos por los centros públicos

«convencionales») moderadas (CREDO, 2009; Zimmer *et al.*, 2009). Destacaríamos aquí el estudio realizado por investigadores de Mathematica (Clark, Gleason, Tuttle, & Silverberg, 2011) sobre el impacto de 36 *charter schools* que en 15 estados distintos de los Estados Unidos. En este estudio se constata cómo, de media, las *charter schools* analizadas no generan ningún impacto significativo (ni positivo ni negativo) en el rendimiento académico global de sus alumnos. Se observan, sin embargo, efectos heterogéneos. Por un lado, se detectan impactos positivos (leves) entre los alumnos desaventajados desde el punto de vista social y académico, y negativos (leves) entre los alumnos más favorecidos. Por otra parte, los impactos tienden a ser más positivos (o menos negativos) en las escuelas que se encuentran en núcleos urbanos o que acogen una composición social más desaventajada (académica y socialmente) que en las escuelas ubicadas en núcleos no urbanos o que acogen una composición más aventajada.

Estudios como los de Angrist *et al.* (2013) y Abdulkadiroglu *et al.* (2009) sobre las *charter schools* de Boston, o como los de Hoxby *et al.* (2009) y Dobbie y Fryer (2011) sobre las *charter* de la ciudad de Nueva York, aportan evidencias sólidas de casos de éxito de estas escuelas. En ambos contextos urbanos, las *charter* producen impactos positivos en el rendimiento y en los niveles de graduación de los alumnos admitidos (en comparación con los resultados de los «perdedores» de la lotería), impactos especialmente relevantes entre los alumnos y centros más desaventajados, y que se incrementan a medida que aumenta la duración de la escolarización en estas escuelas. Más aún, Angrist *et al.* (2013) observan que el beneficio asociado a la asistencia a las *charter schools* de Boston se extiende hasta llegar a incrementar las opciones de acceso a la universidad.

¿Cómo se interpreta el éxito de las *charter* de Boston o de Nueva York? Por una parte, Angrist *et al.* (2011) no niegan la posibilidad de que los diferenciales de efectividad entre las *charter* urbanos y los no urbanos —así como el hecho de que los primeros sean particularmente beneficiosos para los alumnos más desaventajados— puedan explicarse atendiendo a la baja calidad de las escuelas públicas donde muchos de estos alumnos acaban asistiendo al no poder acceder a un *charter* (las escuelas contrafactuales), diferencias de calidad que serían menos marcadas en contextos no

urbanos.⁵ Por otra parte, los autores de las mencionadas evaluaciones coinciden en el argumento de que determinados rasgos característicos de estos centros acaban siendo especialmente relevantes. Buena parte de estos rasgos integran el proyecto educativo denominado No Excuses, suscrito por un sector importante de las *charter schools* urbanos incluidos en estos análisis. Muy brevemente, 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).⁶

En un estudio posterior, el propio Fryer (2011) evaluó los resultados de un experimento consistente en aplicar los rasgos distintivos que parecen hacer funcionar las escuelas *charter* en nueve de los peores centros públicos (por bajo rendimiento) de la ciudad de Houston (en total, más de siete mil alumnos). Pues bien, efectuados los cambios organizativos e incrementados los recursos docentes y materiales necesarios para facilitar la aplicación del nuevo modelo, se comprobó que los alumnos en las escuelas tratadas mejoraron a corto plazo sus niveles de lectura y matemáticas de forma significativa en comparación con los alumnos de los centros de control; más aún, la magnitud del impacto medio de los cambios resultaba casi idéntica a la de los impactos conseguidos por las *charter schools* más efectivos en los estudios antes mencionados de Massachusetts y Nueva York.

A día de hoy, no se tiene evidencia empírica sólida sobre los posibles efectos a largo plazo de los componentes prácticos que el modelo No Excuses ha hecho suyos; en consecuencia, existen dudas sobre su sostenibilidad coste-beneficio (Cullen, Levitt, Robertson, & Sadoff, 2013). Tampoco se ha podido probar la medida en que cada uno de estos componentes contribuye, de forma específica, al impacto global de su aplicación conjunta. No obstante, lo que sí demuestra el estudio de Fryer (2011) es que, *charter* o no *charter*, lo que importa es lo que la escuela hace con su autonomía.

⁵ Clark *et al.* (2011) sostienen este mismo argumento. Si bien los propios autores reconocen que hay que ser cautelosos con los resultados, todo parece indicar que no es que las *charter* urbanas funcionen especialmente mejor que las no urbanas (o las no urbanas especialmente peor que las urbanas); al final, todo depende del nivel de calidad de las escuelas que conforman el escenario contrafactual en uno y otro contexto.

⁶ Como Dobbie y Fryer (2011) recuerdan, la adopción de estos rasgos no es en absoluto privativa del modelo No Excuses; de hecho, estas características hacen referencia a prácticas que la investigación educativa cualitativa lleva estudiando y proponiendo adoptar desde mediados de la década de 1970.

Trayendo el hilo de este razonamiento a la realidad de nuestra red escolar, diríamos que, con un mismo margen de autonomía, lo que los centros hacen o dejan de hacer puede tener un impacto significativo sobre las oportunidades educativas de los estudiantes. Lo que no quita relevancia a la necesidad de evaluar la medida en que la apertura de unos espacios de autonomía escolar u otros (ámbitos curricular, organizativo y de gestión de recursos humanos y materiales) facilita o no prácticas educativas y docentes más o menos efectivas. Eso sí, difícilmente puede avanzarse en esta dirección si no es mediante pruebas piloto aleatorizadas diseñadas para probar el impacto sobre el rendimiento de los alumnos (y sobre otros *outcomes*) de componentes específicos de la autonomía escolar.

4.2. Programas y prácticas de gestión de la diversidad

Nos adentramos, ahora sí, en ámbitos de intervención de carácter más micro. Las prácticas y decisiones aquí seleccionadas no tan solo nos hablan de lo que ocurre en el interior de los centros educativos, sino que todas ellas se plantean de forma explícita como estrategias de gestión de la diversidad, mecanismos que, entre otras cosas, deberían permitir atender de forma efectiva las necesidades educativas del alumnado más vulnerable.

4.2.1. Repetir curso

El número de alumnos de primaria y secundaria que repiten curso en el Estado español es muy superior al correspondiente a la media europea o al de los países de la OCDE (Dupriez, Dumay, & Vause, 2008; Goos *et al.*, 2013), y superior al que nos correspondería a la vista de los niveles competenciales que nuestros alumnos demuestran en las pruebas estandarizadas internacionales (OECD, 2012). En otras palabras, la política de repeticiones representa, en España, una práctica bastante más extendida que en muchos países de nuestro entorno. No han sido pocos los cambios normativos que en el pasado han tratado de regular el uso de esta práctica como mecanismo de gestión de la diversidad en el interior de los centros. Sin ir más lejos, desde la entrada en vigor de la LOGSE, han sido diversos los modelos implicados: de la promoción automática entre cursos (con posibilidad de repetir al finalizar ciclos) a la repetición de cualquier curso a partir de un determinado número de asignaturas

suspendidas.⁷ Sin embargo, ninguno de estos modelos ha sido contrastado empíricamente, y menos aún evaluado en términos de sus posibles impactos en relación con los objetivos sobre los que se pretende incidir (principalmente, mejorar el rendimiento de los alumnos repetidores).

No es fácil llegar a establecer los efectos netos de la repetición. Ciertamente, como acostumbra a ser buena parte de la política de evaluación de alumnos, nos encontramos ante una decisión evaluativa (no promocionar a un alumno) muy idiosincrática; dentro de un mismo sistema educativo, aun existiendo criterios comunes en cuanto a la aplicación de este instrumento, podemos acabar encontrando interpretaciones y usos muy diversos de este criterio, muy variables entre escuelas. Al mismo tiempo, podemos esperar que en los procesos que conducen a la toma de esta decisión (decíamos, procesos idiosincráticos) acaben pesando factores que van más allá del mero rendimiento del alumno –tales como sus actitudes, grado de madurez o nivel de apoyo familiar–, o incluso motivos que poco tienen que ver con lo pedagógico –factores económicos (ahorro de los costes asociados a la repetición) o de imagen (de efectividad del centro y/o de buen perfil de alumnado) –.

En cualquier caso, la evidencia acumulada acostumbra a concluir que la repetición de curso no es únicamente una decisión muy costosa desde el punto de vista económico, sino que además no tiene impactos positivos ni sobre las actitudes ni sobre el progreso educativo de los alumnos repetidores. Revisiones sistemáticas de estudios que han evaluado esta práctica concluyen, antes bien, que la repetición tiende a ser perjudicial, sobre todo para el alumnado más desfavorecido (Holmes, 1989; Jimerson, 2001). El estudio de Jacob y Legren (2009) ofrece un ejemplo interesante de evaluación cuasiexperimental robusta de los impactos de la política de repetición. Los autores aprovechan la introducción en el año 1997, en los centros 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 entre sí los niveles de graduación al finalizar la educación secundaria de dos grupos de alumnos: los que repitieron curso en sexto o en octavo (escuela elemental) obteniendo en el test de competencias una puntuación justo por debajo de la mínima exigida para

⁷ Paralelamente, han sido también diversos y cambiantes los mecanismos dispuestos para recuperar las asignaturas pendientes (recuperaciones durante el curso o en septiembre).

pasar de curso y los que, dentro de las mismas cohortes, promocionaron obteniendo una puntuación justo por encima del umbral mínimo requerido. De esta manera demuestran que el hecho de repetir sexto no afecta a la probabilidad de finalizar con éxito la secundaria; en cambio, repetir octavo sí tiene un impacto negativo sobre esta probabilidad, particularmente entre las chicas afroamericanas.

En España podría aplicarse una estrategia de identificación de este tipo si finalmente acaba instaurándose la reválida externa que plantea la LOMCE en el último curso de la ESO (los resultados de la cual serían vinculantes a los efectos de graduarse o de repetir curso). Alternativamente, podría compararse lo que les sucede a los alumnos que, por una sola asignatura, no superan el número mínimo de materias requerido para pasar de curso con lo que les ocurre a sus compañeros de clase que, por una asignatura, sí superan ese mismo umbral (controlando en todo caso por otras características). Asimismo, a otro nivel, podrían compararse alumnos con el mismo rendimiento académico en cursos consecutivos, unos afectados por la introducción de cambios en los criterios de repetición y otros no (por ejemplo, con el mismo número de asignaturas suspendidas, unos promocionan y otros no). Existen, pues, diversas estrategias para poder evaluar la efectividad de la repetición como mecanismo de gestión de la diversidad, para saber hasta qué punto impulsa o limita el progreso educativo de los alumnos más vulnerables.

4.2.2. Alumnos por aula y agrupaciones de nivel

La ratio de alumnos por clase ha sido y continúa siendo uno de los aspectos de la organización docente y escolar que más atención académica y política ha acaparado en los últimos tiempos. Parte de la centralidad de esta preocupación se debe al hecho de ser este un factor organizativo claramente relacionado con el volumen de recursos de los centros y, por tanto, con la capacidad de inversión pública en educación. Sin ir más lejos, el Ministerio de Educación Cultura y Deporte dictó para el curso 2012-2013 un incremento 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). En efecto, esta medida fue justificada por la necesidad de contención del gasto en educación. En España, los cambios en la regulación de estas ratios han sido frecuentes durante las últimas décadas y en ningún caso han venido respaldados por evidencias empíricas concretas.

Los ejercicios de comparación entre países tienden a evidenciar la existencia de una asociación muy débil entre el número medio de alumnos por aula de los países y su rendimiento académico general (oecd, 2010). Sin embargo, ¿podemos ir más allá y estimar relaciones de causalidad entre medidas de ratio y resultados educativos?⁸ ¿Cómo podríamos hacerlo?

Dos estrategias de evaluación de impacto, ambas de naturaleza cuasiexperimental, merecen una mención especial.⁹ 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, el tamaño de las aulas que los conforman. Así, 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 (por ejemplo, se considera la red escolar de un municipio o de un distrito), 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, la existencia de una normativa estricta sobre el número máximo de alumnos que pueden admitirse por clase ha dado pie, en algunos países, a estudiar los efectos del tamaño de las aulas mediante diseños de regresión discontinua (Angrist

⁸ Las asociaciones que puedan observarse entre estas variables pueden enmascarar situaciones muy diversas. Podemos encontrar, por ejemplo, que las escuelas tiendan a tratar en clases reducidas precisamente al alumnado con peor rendimiento, o que las autoridades educativas opten por ofrecer a los centros desfavorecidos recursos docentes suplementarios, lo que permitiría al conjunto del centro ampliar el número de clases reduciendo su ocupación (hablaríamos aquí de políticas típicamente compensatorias). 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 su composición. Pero podríamos encontrarnos con la situación contraria: las familias con más recursos y más preocupadas por la escolarización de sus hijos, bien escogen aquellos centros que priorizan los grupos-clase reducidos, bien presionan a los centros escogidos para que procedan de esta manera (Schlotter, Schwerdt, & Woessmann, 2010). En ambos casos, la eventual asociación entre clases reducidas y buenos resultados se explicaría por la extracción privilegiada de los alumnos que tenderían a concentrarse en ellas.

⁹ 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 ende, la fiabilidad de sus resultados han sido seriamente cuestionados (Hanushek, 1999).

& Lavy, 1999; Woessmann, 2005a). La aplicación de este diseño se justifica al nos existir 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 permitido 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 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 débiles y poco sostenidos en el tiempo. No obstante, en algunos casos y bajo determinadas circunstancias, esta práctica organizativa sí resulta efectiva. Algunas revisiones sistemáticas de la evidencia acumulada en torno a esta cuestión indican, por ejemplo, que las clases reducidas pueden ser particularmente efectivas para los alumnos más desfavorecidos, con peor rendimiento académico y de menos edad; siempre cuando se trate de reducciones de ratio importantes (orientativamente, por debajo de 15 alumnos por aula) y se cuente con un profesorado preparado para sacarles provecho (McGiverin, Gilman, & Tillitski, 1989; Slavin, 1989; Hattie, 2005).

Junto con el perfil social y académico de los centros educativos, el tamaño de las clases se convierte en un factor claramente condicionante de la capacidad de las escuelas para trabajar con unas fórmulas u otras de agrupación de los alumnos. Sin embargo, no deja de ser cierto que, a igual composición social y a igual ratio de alumnos por clase, los centros educativos aplican estrategias diferentes a la hora de agrupar a los alumnos de un mismo curso a lo largo del año académico. Mientras que algunos fomentan el trabajo en grupos-clase heterogéneos, otros apuestan por estrategias de agrupación homogénea en cuanto a nivel cognitivo. Más aún, estas últimas estrategias varían según su grado de transversalidad (número de horas o asignaturas en las que se trabaja con grupos homogéneos) y de estabilidad (rigidez y perdurabilidad de la agrupación a lo largo del curso). En un extremo encontraríamos la fórmula de los agrupamientos flexibles (grupos reducidos o desdoblamientos de nivel tan solo para algunas asignaturas, siendo posible que los alumnos cambien de asignación grupal a lo largo del curso); en otro extremo, los grupos estables de nivel (clases homogéneas de composición fija durante el curso y, a menudo, con currículums

diferenciados para el conjunto de asignaturas). En España, los centros educativos de primaria y secundaria disponen de plena autonomía a la hora de apostar por unas u otras estrategias. En la práctica nos encontramos con una diversidad manifiesta de modelos, y no son pocos los centros que con el paso de los años han ido introduciendo cambios en su aplicación, probando y ajustando unas fórmulas y otras.

Ahora bien, si nos preguntamos qué evidencia hay acumulada en nuestro país sobre la efectividad de estas fórmulas, de nuevo la respuesta es muy clara: poca. A nivel internacional, por ejemplo, y según muestran los propios datos del pisa, parece claro que los centros que agrupan por niveles tienen tendencia a obtener peores resultados globales que los que no lo hacen. Sin embargo, esta asociación desaparece cuando se tiene cuenta el nivel socioeconómico de los centros y sus alumnos, lo cual, de paso, viene a indicar que la fórmula de los agrupamientos estables de nivel es especialmente prevalente en los centros socioeconómicamente más desfavorecidos.

¿Qué nos dicen los estudios basados en diseños de evaluación más robustos? Si consideramos las revisiones sistemáticas más relevantes sobre la materia, parece claro que el impacto de la agrupación por nivel sobre los logros formativos de los alumnos acostumbra a ser, en el mejor de los casos, modesto (Kulik & Kulik, 1982), cuando no insignificante (Slavin, 1990); y cuando se detectan impactos significativos, estos suelen ser positivos para los alumnos que están en grupos de más nivel y negativos para los alumnos asignados a las clases de nivel más bajo (sobre todo en la educación secundaria) (Schofield, 2010). De forma general, al menos en los países desarrollados,¹⁰ no parece que la fórmula de los grupos de nivel sea, por sí sola, especialmente efectiva como estrategia de incremento de la igualdad de oportunidades en educación, en tanto en cuanto el alumnado socioeconómicamente desfavorecido acostumbra a encontrarse concentrado en los grupos de rendimiento inferior (Hallinan, 2003).

En España, la ausencia de evidencia sobre los efectos de las distintas fórmulas de agrupamiento de los alumnos es manifiesta. Sin embargo, la variación de modelos que podríamos observar (entre centros, entre cursos dentro de los centros, cambios introducidos por los centros a lo largo del tiempo y con reflejo en cursos diferentes en un mismo momento del tiempo, etc.), bien debería ofrecer posibilidades de evaluación

¹⁰ Experimentos realizados en países en desarrollo indican que, en estos contextos, la utilización de grupos reducidos de nivel homogéneo puede resultar beneficioso para el conjunto de alumnos, lo cual se considera una consecuencia del hecho de que los grupos de nivel facilitan la labor docente (Duflo, Dupas, & Kremer, 2011).

robusta de los impactos de las distintas opciones estratégicas. Incluso podrían llegar a plantearse, en el nivel autonómico o local, pilotos experimentales dirigidos a probar la efectividad de unas y otras fórmulas de agrupación de alumnos en distintas asignaturas y cursos.

4.2.3. Innovación y desarrollo pedagógico

Cuando hablamos de innovación y desarrollo pedagógico (maneras de enseñar, estrategias docentes), nos encontramos, por definición, en un terreno enormemente fértil para la producción y acumulación de evidencias, incluso en el caso de aquellos países con poca tradición en la evaluación de efectividad de las políticas educativas, como es nuestro caso. Por un lado, al no encontrarse sujeto a las restricciones normativas que suelen afectar a la articulación de *inputs* de contexto como los tratados anteriormente, este campo aparece mucho más abierto a la experimentación. Por otra parte, en la medida en que el campo propicia el ensayo de innovaciones a pequeña o mediana escala (estrategias docentes específicas, generalmente asociadas a una determinada materia, aprobadas en pocos centros, para un conjunto limitado de clases o secciones), se genera también la necesidad de compartir, de recoger y de acumular evidencias sobre qué funciona y qué no funciona. Es necesario, en todo caso, velar por que ambos procesos —experimentación pedagógica y revisión sistemática— se orquesten de manera que aseguren la producción de evidencias sólidas. Esto quiere decir disponer de buenos experimentos (pruebas piloto de diseño aleatorio que garanticen la validez interna de los resultados obtenidos) y de buenas revisiones sistemáticas (síntesis de evidencias que aporten validez externa al sentido de los resultados de unos experimentos y otros).

De hecho, son numerosos los ámbitos en que ensayo experimental y recopilación de evidencias han sido combinados para intentar aportar un conocimiento sólido y útil para el conjunto del profesorado y los gestores de programas educativos. Esta combinación ha sido trabajada especialmente, una vez más, en los Estados Unidos y, últimamente, también en el Reino Unido. En este terreno, son de especial interés las recopilaciones de síntesis de evidencias que proporcionan los espacios Teaching and Learning Toolkit, de la Education Endowment Foundation (Reino Unido); What Works Clearinghouse, del Institute of Education Sciences (Gobierno federal de los Estados Unidos), y Best Evidence Encyclopedia, de la Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (Estados Unidos). Escapa a

las posibilidades de este texto repasar el conjunto de ámbitos de la práctica pedagógica que han sido objeto de esta estrategia de producción y acumulación de evidencias. Nos conformamos con señalar algunos particularmente relevantes, tanto por el volumen de evidencia que han acumulado (ámbitos prometedores en la lucha contra las desigualdades educativas) como por su proximidad a espacios de intervención que han sido estratégicos y también muy discutidos en nuestro país.

La importancia de las competencias metacognitivas

La defensa de una enseñanza enfocada por competencias y, en particular, hacer énfasis en la importancia de las denominadas *competencias* metacognitivas (motivación, autonomía, flexibilidad, habilidades personales, etc.) son ya lugares comunes del debate curricular en muchas Comunidades Autónomas. No obstante, diversos autores coinciden en afirmar que estos argumentos apenas han trascendido el nivel discursivo, afectando solo superficialmente a la manera en que los contenidos curriculares se ven traducidos en la práctica docente cotidiana, especialmente en la educación secundaria (Coll, 2007). El predominio continúa siendo de los contenidos por encima de las competencias, y de las competencias cognitivas (relacionadas con áreas de saber específicas) por encima de las metacognitivas. Ahora bien, ¿es este un hecho a lamentar? Al fin y al cabo, ¿tenemos pruebas de que unos determinados enfoques curriculares sean mejores que otros? En nuestro entorno no hemos producido tales pruebas, y esto es tanto la causa como la consecuencia de un debate curricular poco interesado en la fundamentación empírica. En otros contextos, por ejemplo, han podido recogerse evidencias sólidas, a menudo experimentales, sobre los niveles de efectividad de programas que promueven el aprendizaje de contenidos mediante la potenciación de competencias relacionadas con la motivación, la autogestión y la capacidad crítica, haciendo hincapié en lo que se ha venido denominar el *aprender a aprender*. De acuerdo con los principales metanálisis sobre la cuestión, existe evidencia sólida de que estos programas tienen la capacidad de incidir positivamente en el progreso académico de los alumnos (Abrami *et al.*, 2008; Dignath, Buettner, & Langfeldt, 2008). El impacto —que es apreciable para el conjunto de materias curriculares y etapas educativas— resulta particularmente significativo entre el alumnado con peor rendimiento, y cuando se trabaja de forma cooperativa en grupos reducidos (Chiu, 1998). En una dirección muy similar apunta la evidencia recogida sobre la efectividad de programas potenciadores de las competencias sociales y

emocionales (Payton *et al.*, 2008; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). Ambos tipos de programas ganan en efectividad cuando son incorporados a la práctica cotidiana de los centros, y cuando cuentan con el apoyo de un profesorado implicado y formado en la materia.

Aprendizaje cooperativo

Se trata de aquel aprendizaje que se adquiere y se moviliza trabajando en grupos reducidos, en torno a actividades y tareas colectivas, y en el que cada alumno desarrolla un papel complementario estratégicamente asignado por el profesorado. En nuestro sistema educativo, esta metodología es habitual en el segundo ciclo de educación infantil y durante los primeros cursos de educación primaria, donde se despliega a través de diferentes procedimientos (tutorías entre iguales, trabajo por rincones o por proyectos grupales). A medida que se avanza en la carrera académica, el uso de estas estrategias va perdiendo peso. Pero, ¿qué garantías empíricas tenemos de que el uso de procedimientos de trabajo cooperativo sea efectivamente positivo para el desarrollo cognitivo (y no cognitivo) de nuestros alumnos? Pues bien, a pesar de la dilatada experiencia de que disponen muchos de nuestras escuelas en la aplicación de estrategias de trabajo cooperativo, a pesar de lo mucho que se ha innovado y se sigue innovando en este terreno y a pesar de la existencia de teorías pedagógicas bien asentadas sobre sus características y sobre lo que habría que esperar de tales estrategias, lo cierto es que no hemos sido capaces de generar suficiente evidencia empírica que justifique (o no) la conveniencia de continuar apostando por ellas. El caso es que estudios (algunos de ellos experimentales) desarrollados en otros contextos (principalmente anglosajones) tienden a atribuir al trabajo cooperativo en el aula impactos significativos positivos sobre el progreso educativo del conjunto de los alumnos en las distintas etapas educativas y materias curriculares. Así lo recogen revisiones sistemáticas de especial relevancia en la materia (Slavin, Cheung, Groff, & Lake, 2008; Slavin & Lake, 2008; Slavin, Lake, Chambers, Cheung, & Davis, 2009; Slavin, Lake, & Groff, 2009). Estas mismas revisiones coinciden en concluir que el aprendizaje cooperativo, para ser efectivo, debe formar parte de una estrategia global bien diseñada, en la que el conjunto de tareas, funciones y recursos que se movilizan disponen de una intencionalidad pedagógica planificada.

*Tutorías 1*1*

Las tutorías 1 por 1 persiguen mejorar el rendimiento competencial de los alumnos más desaventajados a través de un refuerzo individualizado, al margen del grupo-clase ordinario. Los programas de tutorización 1x1 pueden ser diversos, dependiendo, entre otras variables, de si se realizan en horario lectivo o extraescolar, del perfil del tutor (profesores especialistas o no especialistas, paraprofesionales, voluntarios, etc.), de la naturaleza institucional de las entidades promotoras del programa, de los contenidos de las tutorías (más o menos estructurados y focalizados en unas u otras dimensiones competenciales), del perfil y edad del alumnado atendido (curso en que se encuentra, con déficits competenciales más o menos significativos), de la frecuencia y duración de exposición al programa (dosificación), etcétera. En España este sistema de tutorización se encuentra poco implantado como práctica ordinaria de los centros. En cambio, sí se han desarrollado programas de tutorización individual o en grupos reducidos fuera del horario lectivo, dentro o fuera de los centros, y en colaboración con entidades locales, fundaciones o universidades. En Cataluña, por ejemplo, es el caso del programa LEXIT: Lectura per a l'èxit educatiu, 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 cuarto a sexto de primaria;¹¹ el programa Èxit 1: Reforç escolar i activitats complementàries, iniciado en el curso 2001-2002, impulsado por el Consorci d'Educació de Barcelona, y dirigido al alumnado de quinto y sexto de primaria y de primero y segundo de ESO;¹² o los Tallers d'Estudi Assistit, vinculados a los Plans Educatius d'Entorn iniciados por el Departament d'Ensenyament el curso 2004-2005, y dirigidos principalmente al alumnado de secundaria.¹³ Sin embargo, poco sabemos sobre el impacto neto de estos programas, cosa que nos permitiría valorar la conveniencia de potenciarlos y/o probar su traslado al marco ordinario de funcionamiento de los centros. Hay que decir que la evidencia empírica internacional sitúa las fórmulas de tutorización 1x1 como un mecanismo susceptible de generar impactos positivos en el proceso de aprendizaje de los alumnos con peor rendimiento, particularmente durante los primeros cursos de la educación primaria. 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

¹¹ Más información en <http://www.lectura.cat/>

¹² Más información en http://www.edubcn.cat/exit/reforc_escolar_i_activitats_complementaries

¹³ Más información en <http://ves.cat/hrLH>

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 son encargadas a profesorado cualificado o a personal bien formado y asesorado (Slavin, Lake, Davis, & Madden, 2011; Ehri, Dreyer, Flugman, & Gross, 2007).

5. Conclusiones

En último término, el propósito principal de la perspectiva «qué funciona» en educación es promover decisiones y políticas educativas informadas y basadas en la evidencia empírica. La política educativa española, como la de tantos otros países de nuestro entorno, está todavía lejos de este objetivo. 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. Teniendo esto en cuenta, y a la luz de lo que podemos extraer de experiencias y prácticas evaluativas llevadas a cabo en otros contextos, hemos ido situando algunos espacios de oportunidad para el avance en la evaluación de impactos de distintas intervenciones educativas en España, es decir, para conocer mejor en qué medida estas intervenciones consiguen (o no) mejorar las oportunidades educativas de los alumnos.

Sin embargo, si realmente pretendemos avanzar en esta dirección, es importante tener en cuenta dos consideraciones básicas.

En primer lugar, es indispensable poder contar con unos buenos sistemas de información. Toda evaluación de impacto que aspire a ser rigurosa requiere de buenos datos en que basar la construcción de sus variables independientes (características de los alumnos y sus familias, sus clases, profesores, escuelas y entornos, etc.) y de sus variables de outcome (indicadores de rendimiento, de trayectoria educativa, de participación laboral o, incluso, factores menos convencionales relacionados con actitudes y disposiciones hacia la educación o hacia otras instituciones sociales). 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 informaci3n est1 muy limitado o restringido. En nuestro pa3s nos queda mucho camino por recorrer en ambos aspectos.

En segundo lugar, es necesario que las evaluaciones de efectividad sean capaces de llegar al detalle de los impactos, saber qu3 componentes concretos del programa funcionan y cu1les no, averiguar para qu3 colectivos espec3ficos funcionan y para cu1les no, a partir de qu3 momento y durante cu1nto tiempo. Llegar a este nivel de detalle no siempre es f1cil; no siempre se dispone de grupos de control espec3ficos para los distintos niveles de participaci3n en las distintas fases o componentes del programa; no siempre es sencillo dilucidar sobre qu3 subgrupos de participantes tiene sentido analizar la ocurrencia de posibles efectos heterog3neos; no siempre es posible comenzar el proceso de evaluaci3n cuando m1s conviene ni disponer de una ventana temporal lo bastante amplia para la observaci3n de los *outcomes* considerados. Sin embargo, si queremos desentra1ar los mecanismos de efectividad (o ineffectividad) de las intervenciones educativas y saber qu3 hace que un programa funcione o no, hay que impulsar evaluaciones de impacto que apuesten en esta direcci3n. Junto con este requisito, conviene contar con un 1ltimo argumento: si lo que queremos es llegar a saber por qu3 lo que funciona efectivamente funciona —o por qu3 lo que no funciona efectivamente no funciona—, es necesario que podamos complementar el estudio minucioso de impactos con buenas evaluaciones de implementaci3n, susceptibles de capturar la verdadera complejidad de los procesos en juego.

La evaluaci3n deber3a poder responder a ambas preguntas: qu3 funciona y por qu3 funciona, sin perder de vista que solo en la medida en que seamos capaces de responder a la primera pregunta, tendr1 sentido enfrentarnos a la segunda. Esta es la prioridad que ha motivado la elaboraci3n de este art3culo.

6. Referencias

- Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T. J., & Pathak, P. (2009). *Accountability and flexibility in public schools: Evidence from Boston's charters and pilots*. National Bureau of Economic Research. Recuperado a partir de <http://www.nber.org/papers/w15549>
- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102-1134.
- Alegre, M. À., & Ferrer, G. (2010). School regimes and education equity: some insights based on PISA 2006. *British Educational Research Journal*, 36(3), 433-461.
- Angrist, J. D., & 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.
- Angrist, J. D., Pathak, P. A., & Walters, C. R. (2011). *Explaining charter school effectiveness*. National Bureau of Economic Research. Recuperado a partir de <http://www.nber.org/papers/w17332>
- Angrist, J., Cohodes, S.R., Dynarski, S.M., Pathak, P.A., & 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.
- Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic Status, School Quality, and National Economic Development: A Cross-National Analysis of the «Heyneman-Loxley Effect» on Mathematics and Science Achievement. *Comparative education review*, 46(3), 291-312.
- Betts, J. R., & Tang, Y. E. (2011). *The Effect of Charter Schools on Student Achievement*. University of Washington: Center on Reinventing Public Education. http://www.oxydiane.net/IMG/pdf/Charter_NCSRP_BettsTang_Oct11.pdf. Recuperado a partir de http://www.oxydiane.net/IMG/pdf/Charter_NCSRP_BettsTang_Oct11.pdf
- Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final Reading Outcomes of the National Randomized Field Trial of Success for All. *American Educational Research Journal*, 44(3), 701-731.
- Carter, S. C. (2000). *No Excuses: Lessons from 21 High-Performing, High-Poverty Schools*. ERIC.
- Casado, D. (2012). Per què no avaluem les polítiques públiques com els fàrmacs? Una aposta per l'experimentació social. *Avaluació per al Bon Govern*, (3). Recuperado a partir de www.avaluació.cat
- Chiu, C. W. (1998). Synthesizing Metacognitive Interventions: What Training Characteristics Can Improve Reading Performance?.

- Clark, M. A., Gleason, P., Tuttle, C. C., & Silverberg, M. K. (2011). *Do Charter Schools Improve Student Achievement? Evidence from a National Randomized Study*. Mathematica Working Paper. Washington, DC: Mathematica Policy Research. Recuperado a partir de http://www.invalsi.it/invalsi/ri/improving_education/Papers/clark/36.pdf
- Coleman, J. S. (1968). Equality of educational opportunity. *Integrated Education*, 6(5), 19-28.
- Coll, C. (2007). *Currículum i ciutadania: el què i el per a què de l'educació escolar*. Barcelona: Mediterrània.
- Correnti, R. (2009). Examining CSR program effects on student achievement: Causal explanation through examination of implementation rates and student mobility. En *Second Annual Conference of the Society for Research on Educational Effectiveness (Washington)*. Washington, DC. Recuperado 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, CA. Recuperado a partir de http://www.btu.org/sites/default/files/research/credo_standrod_charter_school_performance_full.pdf
- Cullen, J. B., Levitt, S. D., Robertson, E., & Sadoff, S. (2013). What Can Be Done To Improve Struggling High Schools? *Journal of Economic Perspectives*, 27(2), 133-152.
- Di Carlo, M. (2011). The Evidence on Charter Schools and Test Scores. Recuperado a partir de <http://shankerblog.cdjd.info/wp-content/uploads/2011/12/CharterReview.pdf>
- Dignath, C., Buettner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively?: A meta-analysis on self-regulation training programmes. *Educational Research Review*, 3(2), 101-129.
- Dobbie, W., & Fryer Jr, R. G. (2011). *Getting beneath the veil of effective schools: Evidence from New York City*. National Bureau of Economic Research. Recuperado a partir de <http://www.nber.org/papers/w17632>
- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, 3895-3962.
- Duflo, Esther, Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.
- Dumay, X., & Dupriez, V. (2007). Accounting for class effect using the TIMSS 2003 eighth-grade database: Net effect of group composition, net effect of class process, and joint effect. *School Effectiveness and School Improvement*, 18(4), 383-408.

- Dupriez, V., Dumay, X., & Vause, A. (2008). How Do School Systems Manage Pupils' Heterogeneity? *Comparative Education Review*, 52(2), 245-273.
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, 82(1), 405-432.
- Ehri, L. C., Dreyer, L. G., Flugman, B., & 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.
- Fryer, R. G. (2011). *Creating «No Excuses»(Traditional) Public Schools: Preliminary Evidence From an Experiment in Houston*. National Bureau of Economic Research. Recuperado a partir de <http://166.127.16.28/HISDConnectEnglish/Images/Apollo/apollo20whitepaper.pdf>
- Fuchs, T., & Woessmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics*, 32(2-3), 433-464.
- Goos, M., Schreier, B. M., Knipprath, H. M. E., De Fraine, B., Van Damme, J., & 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.
- Hallinan, M. T. (2003). Ability grouping and student learning. *Brookings papers on education policy*, (6), 95-140.
- 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.
- Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, 43(6), 387-425.
- Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. *Flunking grades: Research and policies on retention*, 16, 33.
- 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., & Kang, J. (2009). How New York City's charter schools affect achievement. *The New York City Charter Schools Evaluation Project 2009*.
- Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, 33-58.
- Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, 30(3), 420-437.

- Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, 19(3), 415-428.
- Lupton, R. (2005). Social justice and school improvement: improving the quality of schooling in the poorest neighbourhoods. *British Educational Research Journal*, 31(5), 589-604.
- Luyten, H., Visscher, A., & Witziers, B. (2005). School Effectiveness Research: From a review of the criticism to recommendations for further development. *School Effectiveness and School Improvement*, 16(3), 249-279.
- McGiverin, J., Gilman, D., & Tillitski, C. (1989). A meta-analysis of the relation between class size and achievement. *The Elementary School Journal*, 47-56.
- OECD. (2010). *PISA 2009 Results: What Students Know and Can Do. Vol. 4* (OECD Publishing.). Paris. Recuperado 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. (2012). *Grade Expectations How Marks and Education Policies Shape Students' Ambitions* (OECD Publishing.). Paris. Recuperado a partir de <http://www.oecd.org/pisa/pisaproducts/grade%20expectations%209812091e.pdf>
- Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., & Pachan, M. (2008). The positive impact of social and emotional learning for kindergarten to eighth-grade students. *Chicago, IL: Collaborative for Academic, Social, and Emotional Learning*. Recuperado a partir de <http://esd113.schoolwires.net/cms/lib3/WA01001093/Centricity/Domain/48/casel-fullreport.pdf>
- Rychen, D. S. E., & Salganik, L. H. E. (2003). *Key competencies for a successful life and a well-functioning society*. Hogrefe & Huber Publishers.
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Pergamon Oxford.
- Schlotter, M., Schwerdt, G., & Woessmann, L. (2010). Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide. IZA DP No. 4725. Recuperado 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., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs*. American Educational Research Association Washington, DC. Recuperado a partir de http://69.8.231.237/uploadedFiles/Publications/Books/Estimating_Causal_Effects/ECE_Front-TOC.pdf
- Schofield, J. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. *The Teachers College Record*, 112(5), 1492-1528.

- Schütz, G., West, M., & Woessmann, L. (2007). *School accountability, autonomy, choice, and the equity of student achievement: International evidence from PISA 2003*. Paris: OECD Publishing.
- Slavin, R. E. (1989). Class size and student achievement: Small effects of small classes. *Educational Psychologist*, 24(1), 99-110.
- Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of educational research*, 60(3), 471-499.
- Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective Reading Programs for Middle and High Schools: A Best-Evidence Synthesis. *Reading Research Quarterly*, 43(3), 290-322.
- Slavin, R. E., & Lake, C. (2008). Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis. *Review of Educational Research*, 78(3), 427-515.
- Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective Reading Programs for the Elementary Grades: A Best-Evidence Synthesis. *Review of Educational Research*, 79(4), 1391-1466.
- Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, 6(1), 1-26.
- Slavin, R. E., Lake, C., & Groff, C. (2009). Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis. *Review of Educational Research*, 79(2), 839-911.
- Teddlie, C., & Reynolds, D. (2000). *The international handbook on school effectiveness research*. Psychology Press.
- Thernstrom, A., & Thernstrom, S. (2004). *No excuses: Closing the racial gap in learning*. Simon and Schuster.
- Thrupp, M. (1999). *Schools Making a Difference: Let's Be Realistic! School Mix, School Effectiveness and the Social Limits of Reform*. ERIC.
- Thrupp, M. (2001). Sociological and political concerns about school effectiveness research: Time for a new research agenda. *School Effectiveness and School Improvement*, 12(1), 7-40.
- Whitman, D. (2008). *Sweating the small stuff: Inner-city schools and the new paternalism*. Thomas B. Fordham Institute Washington, DC.
- 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., & Schuetz, G. (2009). *School accountability, autonomy and choice around the world*. Edward Elgar.

Woessmann, L., & 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.

Zimmer, R. W., Gill, B., Booker, K., Lavertu, S., Sass, T. R., & Witte, J. (2009). *Charter schools in eight states: Effects on achievement, attainment, integration and competition* (Vol. 869). RAND Corporation.