LA EVALUACIÓN EN EL CONTEXTO EDUCATIVO

José Manuel CORDERO FERRERA
Francisco PEDRAJA CHAPARRO
Rosa SIMANCAS RODRÍGUEZ
Universidad de Extremadura

Resumen

El presente estudio ofrece un breve repaso de la literatura nacional e internacional centrada en la evaluación de factores vinculados con el rendimiento escolar, prestando especial atención a la descripción de los métodos empleados para realizar dichas evaluaciones y los requisitos de información que requiere su correcta implementación. Asimismo, se presenta un ejemplo práctico de evaluación educativa realizada en la región de Extremadura con el propósito de mejorar la formación de su población a través de un incentivo monetario.

Palabras clave: evaluación, educación, inferencia causal, políticas educativas.

Abstract

This study offers a brief literature review of national and international studies focused on the evaluation of factors related to student achievement, paying special attention to the description of the methods used to conduct such evaluations and the information requirements for their correct implementation. We also present an illustrative example of an educational evaluation conducted in the region of Extremadura with the aim of improving the education level of the population through a monetary incentive.

Keywords: evaluation, education, causal inference, education policies.

JEL classification: I21, H52, C18.

I. INTRODUCCIÓN

JUSTIFICAR la evaluación en educación puede resultar algo obvio. El gasto en educación constituye uno de los programas fundamentales del Estado de bienestar y tiene efectos positivos en los dos objetivos básicos de la economía pública, la eficiencia (crecimiento) y la equidad (igualdad de oportunidades), además de otras consecuencias favorables para la sociedad como son la mejora en los niveles de salud, la reducción de la actividad delictiva o la mejora de su calidad democrática. Por su parte, la evaluación de las políticas públicas constituye un elemento de legitimación de la actuación del sector público en la economía y forma parte de la rendición de cuentas de los Gobiernos a sus ciudadanos. Desde una perspectiva más pragmática, la evaluación constituye un elemento fundamental en la mejora de la gestión pública al basar los programas públicos en función de sus resultados más que en buenas intenciones o meras intuiciones.

Dada la amplitud del contenido, centramos nuestro artículo en la descripción de trabajos de evaluación relacionados con factores explicativos del rendimiento escolar. El interés de la selección se justifica por la relación de estos resultados con el crecimiento económico. Según Hanushek y Woessman (2011), medidas cualitativas de capital humano (resultados escolares en pruebas internacionales estandarizadas) explican una parte importante de la variación interanual en las tasas de crecimiento del PIB entre países (1).

Los economistas analizamos esta cuestión mediante una función de producción educativa, cuyo análisis empírico se ha visto impulsado y enriquecido tanto por la disponibilidad de datos derivados de la generalización de evaluaciones internacionales de los sistemas educativos como por el desarrollo de métodos econométricos que permiten analizar la amplia variedad de factores que pueden influir en los resultados educativos, que van desde el entorno personal y familiar de los estudiantes y el aula, hasta los recursos escolares y factores de tipo institucional (Todd y Wolpin, 2003; Creemers y Kyriakides, 2007). Dado el espacio del que disponemos, hacemos un repaso ejemplificativo y no exhaustivo de estudios que utilizan técnicas de evaluación de impacto en el análisis de la función de producción educativa desde una doble perspectiva, internacional y española. Continuamos el trabajo con un ejemplo de evaluación educativa, el Programa 18-25, puesto en marcha en Extremadura con la intención de mejorar su capital humano con la obtención de la Enseñanza Secundaria Obligatoria (ESO) ofreciendo un incentivo monetario. Por último, cerramos el artículo con el habitual apartado de conclusiones.

II. LA EVALUACIÓN EN EDUCACIÓN EN PERSPECTIVA INTERNACIONAL

El interés de los economistas por la función de producción educativa tiene su origen en las polémicas conclusiones obtenidas en el famoso informe Coleman (Coleman *et al.,* 1966) cuyo objetivo era comprobar la importancia de los recursos escolares en el éxito escolar, principalmente, en los grupos socioeconómicos más desfavorecidos. Esta investigación, basada en datos de más de 700.000 estudiantes en Estados Unidos, llegó a la conclusión de que los recursos escolares explicaban una parte muy reducida de las diferencias registradas en los resultados escolares en comparación con la importancia que tenían los factores socioeconómicos. Posteriormente, Hanushek (1986), en otro de los trabajos fundamentales dedicados a la función de producción educativa (2), llegó a una conclusión similar, destacando la escasa importancia de ese tipo de recursos.

En buena medida, esos resultados tenían su origen tanto en los datos disponibles como en las técnicas econométricas utilizadas que impedían tener en cuenta la diversidad de factores que forman parte del proceso educativo y la complejidad de sus relaciones. Diversidad y complejidad que dificultaban la identificación de relaciones causales entre las variables analizadas y con ello la posibilidad de evaluar el impacto de las intervenciones en el ámbito educativo.

Para poder evaluar correctamente esas intervenciones es necesario corregir los posibles sesgos que provoca la existencia de endogeneidad en los datos, derivada de la omisión o imposibilidad de observación de factores relevantes, la existencia de causalidad inversa, el sesgo de autoselección o los errores de medición en las variables de interés observadas. La búsqueda de soluciones para evitar los problemas vinculados con la endogeneidad se ha convertido en una de las principales preocupaciones de los investigadores que trabajan la economía de la educación (3). Nuestro propósito en este trabajo no es analizar las técnicas existentes para corregir esos problemas, sino mostrar aquellas que se han utilizado más habitualmente en la literatura de evaluación de políticas educativas, destacando algunas de sus ventajas e inconvenientes, así como los requisitos que exige su correcta implementación.

1. La estrategia de la aleatorización

La mejor manera de corregir los posibles sesgos de selección e identificar los efectos causales es mediante la realización de un experimento aleatorio, es decir, seleccionando aleatoriamente una muestra de la población y, dentro de ella, asignando al azar a los individuos a los grupos de tratamiento y control. De este modo se crean dos grupos con características estadísticamente equivalentes salvo en lo referido a su participación en el programa objeto de evaluación. Esto permitirá concluir que las diferencias encontradas entre los resultados medios del grupo de individuos tratado y del grupo de control se deben al tratamiento (programa), ya que otras posibles explicaciones quedan anuladas por la propia estrategia del proceso de asignación aleatoria (Murname y Willett, 2011).

Uno de los primeros experimentos aleatorios en el ámbito educativo fue el Proyecto Perry, un programa que se puso en marcha en los años sesenta en Estados Unidos con el propósito de analizar la efectividad de los programas de educación preescolar. Este estudio contó con la participación de 123 alumnos en riesgo de fracaso escolar, a los que se dividió aleatoriamente en dos grupos; uno siguió un programa preescolar de alta calidad basado en el enfoque de aprendizaje activo; y otro, que actuó como grupo de control, no tuvo acceso al programa. La principal particularidad de este estudio es que se hizo un seguimiento longitudinal de los alumnos. Han sido muchos los trabajos que han explotado esta base de datos, aunque los análisis más relevantes, aquellos realizados por el premio nobel de economía James Heckman y sus colaboradores, concluyeron que los beneficiarios del programa obtuvieron mejoras significativas en los ámbitos educativo y laboral (acceso al empleo, estabilidad laboral y salarios), así como mejores niveles de salud y menores tasas de criminalidad (Heckman *et al.,* 2010a; 2010b).

Otro famoso experimento en este campo es el Proyecto STAR (Student/Teacher Achievement Ratio), implementado en los años ochenta en el estado de Tennessee (Estados Unidos) cuyo objetivo fue comprobar el efecto del tamaño de la clase en los resultados escolares y en el que participaron alrededor de 11.600 alumnos. Se trata de un experimento controlado que comenzando en la guardería se extendía hasta tercero de primaria y que asignaba aleatoriamente a los alumnos en clases de tamaño reducido (entre 13 y 17 alumnos) y en

clases de tamaño normal (entre 22 y 25) en las que un ayudante asistía al profesor (Word *et al.,* 1990). Los trabajos que analizaron los resultados del experimento demostraron que aquellos alumnos que asistieron a las clases reducidas consiguieron mejores resultados y tuvieron mayor probabilidad de asistir a la universidad, siendo los efectos mayores en el caso de alumnos pertenecientes a entornos más desfavorecidos (Finn y Achilles, 1999; Chetty *et al.,* 2011).

Con el paso del tiempo, este tipo de experimentos se ha convertido en una práctica más frecuente en la evaluación de las políticas educativas, especialmente en los países anglosajones, donde la cultura de la evaluación está mucho más arraigada. No obstante, también es posible encontrar la aplicación de estas metodologías en países en vías de desarrollo. Los trabajos de Kremer (2003) y Duflo y Kremer (2005) resumen algunas evaluaciones de diferentes intervenciones educativas como son la entrega de dinero para promover la asistencia a la escuela, las comidas gratuitas en centros escolares, la gratuidad de uniformes escolares, la utilización de libros de texto o la incorporación de un segundo maestro a las clases.

Lamentablemente, la realización de estos experimentos se ve limitada por inconvenientes de distinta naturaleza, desde los económicos (por su elevado coste) hasta los éticos (selección de los grupos), pasando por la imposibilidad de su aplicación dado el carácter general de las políticas públicas. En todo caso, parece preferible ensayar las políticas públicas sobre las que existe un alto grado de incertidumbre mediante un programa piloto con el que evaluemos su impacto que aplicar esa política de forma generalizada al conjunto de la población.

Por otra parte, muchas políticas educativas no se diseñan con la intención de ser evaluadas. En este último caso, los investigadores deberán trabajar con los datos observados una vez que la política educativa ha sido implementada, tratando de imitar las condiciones de un experimento aleatorio mediante la utilización de técnicas cuasiexperimentales.

2. Aproximaciones cuasiexperimentales

Estas aproximaciones exigen controlar por todas las diferencias (sean o no observables) existentes entre los grupos de tratamiento y de control para poder estimar correctamente el efecto del tratamiento. Entre ellas cabe destacar el método de diferencias en diferencias (DiD) y sus variantes (efectos fijos y control sintético), la regresión en discontinuidad (RD) y el enfoque de variables instrumentales (VI) (4)(5). La elección entre estos métodos dependerá de las características del programa a evaluar y del tipo y calidad de los datos disponibles.

La utilización de la estrategia de *diferencias en diferencias* requiere disponer de información sobre los individuos que componen los grupos de tratamiento y control antes y después de la implementación de una determinada intervención, lo que exige la utilización de datos longitudinales. Esto explica que gran parte de los estudios empíricos que aplican esta técnica se hayan llevado a cabo en países desarrollados, donde este tipo de datos suelen estar disponibles gracias a la realización periódica de evaluaciones nacionales a gran escala o donde existen registros administrativos de carácter censal (6).

Un ejemplo puede encontrarse en el trabajo de Machin y McNally (2008) que aplica la técnica de DiD para evaluar el impacto de una hora semanal adicional de lectura introducida en las escuelas de enseñanza primaria de Inglaterra a finales de los años noventa, llegando a la conclusión de que tuvo un impacto positivo. También existen varios trabajos que utilizan esta metodología con datos administrativos para evaluar el impacto de programas de incentivos al profesorado aplicados en diferentes estados americanos (Eberts *et al.,* 2002; Imberman y Lovenheim, 2015) que, en su mayoría, no parecen contribuir significativamente a la mejora del rendimiento de los estudiantes (7). Afortunadamente, en las dos últimas décadas, cada vez hay más países en vías de desarrollo que han hecho el esfuerzo de recopilar microdatos sobre los estudiantes y las escuelas de manera periódica, lo que permite la construcción de bases de datos con estructura de panel y facilita la evaluación de diferentes políticas educativas aplicando la técnica de DiD. Los trabajos de Lai *et al.* (2011) y Ding *et al.* (2020) para China, Angrist *et al.* (2006) o Laajaj *et al.* (2022) con datos colombianos o Glewwe *et al.* (2009) y Lucas y Mbiti (2012) para Kenia son buenos ejemplos.

A pesar de que uno de los requisitos del enfoque DiD es poder observar a los individuos antes y después de la implantación de una determinada política, también es posible aplicarlo con datos de sección cruzada (un único período) siempre que se disponga de, al menos, dos medidas diferentes de la variable de interés para cada individuo de

los grupos de tratamiento y control. Jürges *et al.* (2005), por ejemplo, analizan el efecto causal de la implantación de pruebas estandarizadas de conocimientos en Alemania aprovechando el hecho de que en algunos Estados existen este tipo de pruebas solamente para las matemáticas y no para las ciencias. Por tanto, la primera diferencia es la registrada entre las dos materias y la segunda es la diferencia entre estudiantes en Estados con y sin pruebas. Schwerdt y Wuppermann (2011) y Bietenbeck (2014), por su parte, también analizan el efecto de dos tipologías de prácticas docentes (modernas frente a tradicionales), explotando las diferencias existentes entre los resultados obtenidos en las dos competencias evaluadas en el estudio TIMSS para controlar los rasgos no observados de los estudiantes.

La principal crítica que puede hacerse a estos trabajos es que la variación existente entre los resultados obtenidos por los alumnos en las diferentes materias evaluadas en las pruebas internacionales puede ser ficticia. Esto se debe a que las medidas representativas de esos resultados, los denominados valores plausibles, se generan mediante un proceso de imputación que toma como referencia las respuestas de los estudiantes a un número reducido de preguntas y a su contexto personal y familiar (8). Por tanto, es posible que la variación existente entre asignaturas se explique únicamente por el proceso de imputación utilizado en estas pruebas, especialmente aquellos casos en los que los estudiantes solo hayan respondido a preguntas de una única materia (Jerrim *et al.,* 2017).

La aplicación de la *regresión en discontinuidad* también está condicionada en gran medida por la disponibilidad de un volumen de datos suficiente alrededor del punto de corte en la variable que determina quiénes podrán beneficiarse del programa y quiénes formarán parte del grupo de control. Por esa razón, no es extraño que muchos de los trabajos que emplean este enfoque metodológico utilicen datos administrativos. Este tipo de diseño fue utilizado por primera vez para analizar el efecto de un programa de becas al que accedían aquellos estudiantes que obtuvieran una nota superior a un valor obtenido mediante un test específico diseñado para el programa (Thistlewaite y Campbell, 1960).

Desde entonces, se ha utilizado para analizar políticas educativas muy diversas (9). Esta técnica ha sido profusamente utilizada para analizar el efecto del tamaño de clase, aprovechando la existencia de normas en muchos países que fijan un tamaño máximo de clase en las escuelas y que obligan a su división en grupos más reducidos en caso de superar ese máximo. En un trabajo reciente, Angrist *et al.* (2019) aplicaron el método de RD a una base de datos de escuelas públicas de Israel y no identificaron ningún efecto significativo de la política de reducción del tamaño de la clase (10). En otro relevante trabajo, Jacob y Lefgren (2009) utilizan esta técnica para analizar el efecto de la política de repetición de curso a partir de datos administrativos de las escuelas públicas de Chicago, llegando a la conclusión de que obligar a repetir a estudiantes con bajo rendimiento en la escuela primaria aumenta sustancialmente la probabilidad de que abandonen la escuela secundaria.

El uso del método de *variables instrumentales* tiene el propósito de eliminar, o al menos mitigar, los problemas de sesgo relacionados con la omisión de variables relevantes o la posible causalidad inversa. Esta alternativa no es tan exigente con los datos necesarios para su aplicación, lo que resulta clave es identificar el instrumento apropiado que pueda ser considerado como una alteración exógena del proceso educativo (11).

Un hecho que puede facilitar la identificación del instrumento es la existencia de una norma o un cambio en la regulación del sistema educativo. Por ejemplo, Machin *et al.* (2007) aprovechan un cambio en las reglas de asignación de recursos para la adquisición de tecnologías de información y conocimiento (TIC) en los distintos distritos escolares de Inglaterra como instrumento para evaluar el efecto causal del gasto en TIC sobre los resultados de los alumnos, identificando un efecto positivo en dos de las tres competencias evaluadas en su estudio.

Otra variable frecuentemente utilizada como instrumento es la fecha de nacimiento de los alumnos, en combinación con la de su incorporación a la escuela, para corregir posibles sesgos asociados por la no consideración de las capacidades innatas de los estudiantes en el análisis. Esta estrategia se ha empleado tanto para analizar el efecto de los años de escolarización (Angrist y Krueger, 1991) como las consecuencias de la política de repetición (Green y Winters, 2007) o simplemente para identificar el efecto de la edad (o el grado de madurez) de los alumnos sobre su rendimiento académico (Lee y Fish 2010).

Por último, una estrategia de identificación habitual consiste en utilizar la localización de las escuelas o la distancia de los alumnos a ellas para estimar su efecto sobre el rendimiento. En particular, este enfoque ha sido principalmente empleado en trabajos que analizan si los alumnos que asisten a escuelas privadas presentan resultados significativa y estadísticamente diferentes de aquellos matriculados en centros de titularidad pública (Vandenberghe y Robin, 2004; Pfeffermann y Landsman, 2011).

En la literatura más reciente podemos encontrar un conjunto de trabajos que utilizan la información que proporcionan profesores y directores de los centros educativos como instrumento para corregir la posible endogeneidad existente en los datos de los estudiantes al evaluar el efecto de una determinada política sobre sus resultados académicos. Por ejemplo, al estudiar los efectos del tiempo dedicado a los deberes, Gustafsson (2013) emplea la información proporcionada por los profesores como instrumento del tiempo que declaran los alumnos dedicar a sus tareas. Sin embargo, el uso de estos instrumentos y, por tanto, los resultados que se derivan de ellos resultan bastante más cuestionables por la dificultad de otorgar un carácter exógeno a las opiniones personales.

III. LA EVALUACIÓN EN EDUCACIÓN EN ESPAÑA

Hasta principios de este siglo, apenas existían en nuestro país trabajos empíricos que analizaran los factores explicativos del rendimiento escolar de los estudiantes debido, en gran medida, a la escasez de información adecuada para realizar este tipo de estudios. Los escasos ejemplos que había estaban, en su mayoría, basados en datos agregados referidos a las escuelas, lo que limitaba el alcance de sus resultados. Sin embargo, desde que España comenzó a participar en las principales evaluaciones internacionales (TIMSS, PIRLS, PISA O PIAAC), se produjo un importante avance en esta línea de investigación gracias a la disponibilidad de una amplia variedad de microdatos sobre múltiples factores que pueden influir en el proceso educativo. Asimismo, la realización de dos evaluaciones nacionales a gran escala (Evaluaciones generales de diagnóstico –EGD–), impulsadas por el Instituto Nacional de Evaluación Educativa (INEE) en primaria (año 2009) y secundaria (año 2010), ofreció a los investigadores dos fuentes de información adicionales para el estudio específico de nuestro sistema educativo.

Al igual que en el contexto internacional, los primeros trabajos que exploraron los determinantes del rendimiento académico se centraron en la identificación de relaciones estadísticamente significativas entre variables (12). Sin embargo, siguiendo la tendencia descrita en el ámbito internacional, pronto comenzaron a surgir trabajos en los que se utilizaban modelos de inferencia causal con el propósito de aproximar las condiciones ideales de los experimentos aleatorios a pesar de que, como se ha mencionado anteriormente, estas bases de datos no están diseñadas para la aplicación de esas técnicas.

Uno de los estudios pioneros en este sentido fue el de Salinas y Santín (2012), en el que se empleaba el uso de variables instrumentales para corregir el posible sesgo en la estimación de la función de producción educativa al comparar el rendimiento de los alumnos inmigrantes y nacionales que asistían a escuelas públicas y concertadas. Una de sus principales conclusiones fue que la concentración de alumnos inmigrantes en centros públicos repercutía negativamente sobre los nativos, mientras que una baja concentración de aquel tipo de alumnado en las escuelas concertadas mejoraba el rendimiento educativo de los mismos. A una conclusión similar llegaron Pedraja *et al.* (2016) al analizar el efecto de la concentración de inmigrantes en determinados centros educativos utilizando datos procedentes de la base de datos PISA, aunque en su caso la estrategia de estimación se basaba en el uso de diferencias en diferencias.

El efecto de la política de repetición de curso sobre el rendimiento escolar de los alumnos es otra de las cuestiones que ha atraído la atención de los investigadores españoles debido a que nuestro país alcanza una de las mayores tasas de repetición de Europa (López-Rupérez *et al.,* 2021). La obtención de resultados fiables exige tener en cuenta el problema de causalidad inversa, ya que si los repetidores presentan peores resultados puede ser debido a la repetición o a que ya tenían un peor rendimiento antes de repetir. Los únicos estudios empíricos que han empleado enfoques metodológicos que permiten corregir posibles sesgos derivados de este problema han sido los de García-Pérez *et al.* (2014), mediante el uso de variables instrumentales, y Choi *et al.* (2018), incorporando el rendimiento previo de los alumnos mediante la construcción de un pseudopanel. Ambos llegan a la conclusión de que la estrategia de repetición tiene un efecto negativo sobre el rendimiento, siendo este mucho más grave en primaria que en secundaria.

También podemos encontrar estudios que han tratado de establecer una relación causal entre la asistencia a la educación preescolar y los resultados educativos, intentando corregir posibles sesgos provocados por la endogeneidad implícita de esta decisión (los años de asistencia a educación infantil están correlacionados con variables no observables como sucede con la motivación de los padres), mediante diferentes aproximaciones cuasiexperimentales. En todos ellos se llega a la conclusión de que existe una relación estadísticamente significativa y positiva entre ambas variables (Santín y Sicilia, 2015; Mancebón *et al.,* 2018). Asimismo, Felfe *et al.* (2015) analizaron el impacto que tuvo la ampliación gratuita de la educación infantil a los tres años con la implantación de la LOGSE (Ley de Ordenación General del Sistema Educativo) a principios de los años noventa, explotando el experimento natural que se produjo como consecuencia de la distinta extensión de la reforma en las comunidades autónomas. Utilizando datos de diferentes oleadas de PISA y un enfoque de diferencias en diferencias, se comprobó que la reforma afectó positivamente a la promoción de curso durante la educación primaria y supuso una mejora en los resultados educativos en lectura y matemáticas que se mantuvo en el largo plazo. Además, este efecto fue mucho más acusado en los alumnos con peor nivel socioeconómico.

También contamos con estudios que han explotado la información disponible en las bases de datos internacionales y nacionales para analizar el efecto de diferentes factores como el número de horas de clase impartidas o las estrategias docentes que emplean los profesores. Sobre el primer aspecto, López-Agudo y Marcenaro (2019a, 2019b), tras aplicar un modelo de efectos fijos a los datos disponibles en PISA, obtienen que el tiempo de instrucción semanal no parece afectar al rendimiento académico. En cuanto al segundo, la efectividad de métodos de enseñanza más modernos frente a los tradicionales, la evidencia no es concluyente ya que mientras Hidalgo y López-Mayán (2018), con datos procedentes de la EGD y aplicando un enfoque DiD, encuentran un efecto positivo sobre el rendimiento académico, Cordero y Gil-Izquierdo (2018), con datos procedentes de PISA y aplicando variables instrumentales, obtienen un efecto negativo. Sin embargo, ambos estudios coinciden en la identificación de un efecto positivo en el uso de las estrategias docentes tradicionales.

Un aspecto que ha suscitado gran interés acrecentado por la actual pandemia, ha sido el impacto de la incorporación de las TIC en las escuelas a través de diferentes programas impulsados tanto por el Ministerio de Educación como por las comunidades autónomas. Lamentablemente, la mayoría de estas iniciativas no obedecieron a un plan consensuado para todo el territorio nacional ni establecieron un criterio de evaluación rigurosa, por lo que resulta muy difícil comprobar si resultaron ser eficaces. La excepción ha sido el programa Escuela 2.0, cuyo objetivo era mejorar la dotación de ordenadores de los centros educativos y para el que existen estudios que han analizado su impacto mediante la aplicación de técnicas de inferencia causal a los datos disponibles en PISA (Villaplana, 2014; Cabras y Tena, 2016). Ambos concluyen que el efecto fue moderadamente positivo solamente en el caso de que tal política viniera acompañada con cambios en la metodología docente del profesorado.

Al margen de los trabajos mencionados, todos ellos basados en información procedente de bases de datos internacionales, existen otros que han explotado la información de la que disponen las comunidades autónomas gracias a las evaluaciones que estas realizan periódicamente en su territorio. Su carácter censal y estructura longitudinal se adapta mucho mejor a la utilización de técnicas de inferencia causal. Así, podemos encontrar estudios que explotan los datos administrativos de los estudiantes de Andalucía para analizar cómo afecta al rendimiento escolar el tamaño de clase (López-Agudo y Marcenaro, 2021) o el tiempo que dedican a la realización de deberes y a la lectura (Jerrim *et al.,* 2019; 2020). También se pueden encontrar trabajos que analizan los datos censales de los estudiantes madrileños para examinar la efectividad de los programas de bilingüismo implantados en esa Comunidad (Anghel *et al.,* 2016; Mediavilla *et al.,* 2019) o los datos relativos a estudiantes catalanes para evaluar el impacto del Programa para la Mejora de la Calidad Educativa implantado en varios centros educativos de Cataluña (López-Torres *et al.,* 2019). Por último, los microdatos de los estudiantes de primaria y secundaria del País Vasco han sido utilizados como referencia para analizar la pérdida de aprendizaje provocada por el cierre de los colegios por la pandemia de la COVID-19 (Arenas y Gortazar, 2022).

Lamentablemente, estas bases de datos son de carácter informativo e interno y no están disponibles al público en general, de manera que su explotación se encuentra supeditada a la voluntad de que la institución responsable de los datos los

facilite a los investigadores, lo que solo sucede de forma excepcional. Esta situación debería corregirse pues son muchos los recursos destinados a recopilar esa información que debería estar disponible de modo general. Como dijimos, la evaluación de las políticas constituye un elemento de legitimación de la actuación del sector público en la economía y forma parte de la rendición de cuentas de los Gobiernos a sus ciudadanos.

La gran asignatura pendiente en nuestro país es el desarrollo de experimentos aleatorios para llevar a cabo la evaluación de las políticas educativas. Es cierto que existen algunas excepciones dignas de mención, como el estudio piloto desarrollado por el Banco de España para evaluar la efectividad de la implantación de un programa de educación financiera (Bover *et al.,* 2018) o el proyecto financiado por el programa Erasmus+ de la Unión Europea para medir hasta qué punto la educación cívica activa influye en la participación y compromiso democrático de los jóvenes mediante una intervención experimental a gran escala en la que participaron Francia, Inglaterra, Grecia y España (Briole *et al.,* 2022). Por otro lado, cabe destacar la labor llevada a cabo en los últimos años por la Fundación «la Caixa», que se ha decidido a promocionar la innovación educativa basada en evidencias poniendo en marcha el programa «*Evaluación de programas educativos*» de EduCaixa, cuyo objetivo es fomentar la evaluación de intervenciones e innovaciones educativas poniendo en contacto centros educativos con propuestas educativas innovadoras y equipos de investigación (13).

IV. UN EJEMPLO DE EVALUACIÓN EDUCATIVA EN ESPAÑA: EL PROGRAMA 18-25

En este apartado se ofrece un breve resumen de un estudio empírico realizado en el contexto español con el propósito de mejorar el nivel de formación de la población desempleada. Se trata de una iniciativa puesta en marcha en Extremadura, donde en el año 2012 el 41 por 100 de los desempleados inscritos en el Servicio Extremeño Público de Empleo carecían de la Educación Secundaria Obligatoria (ESO). Con el propósito de reducir esas cifras, el Gobierno autonómico aprobó, en noviembre de ese mismo año, el denominado Programa 18-25, el cual se mantuvo vigente durante tres cursos académicos, suprimiéndose con el cambio de Gobierno regional en mayo de 2015.

El objetivo de este programa era reducir, mediante un incentivo económico de 1.000 euros, el número de desempleados extremeños que carecían de la formación básica obligatoria y con ello mejorar sus posibilidades de incorporación al mundo laboral. Los beneficiarios potenciales eran todos los desempleados con edades comprendidas entre 18 y 25 años que no contasen con el título de la ESO. Además, en el caso de las mujeres, también podían beneficiarse las mayores de 25 años que se encontrasen en situación de desempleo de larga duración. El programa se gestionaba de acuerdo con lo establecido en la Educación Secundaria para Adultos (ESPA) en Extremadura y, por tanto, estaba organizado en módulos en lugar de asignaturas. El alumno interesado en acogerse al Programa 18-25 debía indicarlo expresamente en su solicitud de admisión en formación de adultos, matriculándose del total de módulos pendientes hasta un máximo de seis (que era el número máximo de módulos permitidos por curso académico en la ESPA). El incentivo económico estaba asociado al rendimiento académico, de manera que los participantes en el programa debían asistir regularmente a clase y aprobar los módulos en los que estuvieran matriculados. Así, el alumno que cumpliera los requisitos anteriores recibiría un pago de 500 euros al finalizar cada cuatrimestre.

Con esta evaluación se pretende comprobar si el programa incrementó la probabilidad de obtener el título de formación básica. Para ello, compararemos el grado de éxito de aquellos que se acogieron al mismo con los que, asistiendo durante el mismo curso a la ESPA, no pudieron beneficiarse del programa por no cumplir el requisito de la edad (ser mayores de 25 años). Este análisis se centrará en la segunda convocatoria del programa, la correspondiente al curso 2013-2014 y, puesto que no es posible evaluar el efecto del programa sobre toda la población objetivo por tener requisitos diferentes según género, optamos por observar su impacto únicamente sobre la población masculina.

Dentro de las diferentes técnicas cuasiexperimentales, se seleccionó como herramienta de análisis la regresión en discontinuidad por ser la que mejor se ajustaba a la naturaleza del programa que, al impedir el acceso a los mayores de 25 años, provoca una discontinuidad en el tratamiento. Este enfoque permite comparar a individuos en un entorno cercano al punto de corte con características razonablemente similares, lo que facilita la identificación del efecto causal de la política.

Las regresiones en discontinuidad se aplican en aquellas situaciones en las que los beneficiarios y no beneficiarios de una política específica son determinados por el hecho de encontrarse por encima o por debajo de un valor concreto $\bar{x}$ de una variable de asignación X_i. La principal ventaja de las regresiones en discontinuidad es que, al comparar los resultados obtenidos por unidades situadas en un entorno cercano (por encima y por debajo) del punto de corte, nos acerca a lo conocido como asignación aleatoria al tratamiento y, por tanto, cualquier diferencia en los resultados puede interpretarse como efecto causal del programa (Gertler *et al.,* 2016). La regresión en discontinuidad puede ser estricta (cuando la variable de asignación determina los grupos de tratamiento y control de forma inequívoca) o difusa (la variable de asignación no permite determinar con exactitud la participación en el programa, sino más bien la probabilidad de participar en el mismo [Schlotter *et al.,* 2011]). Esto último sucede en nuestro caso, donde existen individuos que, cumpliendo los requisitos, decidieron no participar en el programa mientras que otros individuos, aun no cumpliendo el requisito de la edad, acabaron formando parte del grupo de tratados. Esto significa que la variable de asignación (Edad) no determinó exactamente la participación, lo que nos llevó a estimar una regresión discontinua difusa.

Las regresiones discontinuas difusas pueden interpretarse desde un enfoque de variables instrumentales, donde, a partir de la variable de asignación X_i, se genera una variable I_i, que funciona como instrumento de la variable tratamiento D_i (Angrist y Pischke, 2014). De este modo, las ecuaciones estimadas son las siguientes:

1.ª etapa o ecuación de tratamiento:

$$D_i = \gamma_0 + \gamma_1 I_i + \gamma_2 X_i + \gamma_3 Z_i + \varepsilon_1 \qquad (1)$$

2.ª etapa o ecuación de resultado:

$$Y_i = \beta_0 + \beta_1 \widehat{D}_i + \beta_2 X_i + \beta_3 Z_i + \varepsilon_2 \qquad (2)$$

donde Y_i representa la medida del impacto del programa (en nuestro caso, toma valor 1 si, al finalizar el curso académico, el individuo consiguió el título de la ESO y 0 en caso contrario); D_i indica el tratamiento real (variable dicotómica que toma valor 1 si el alumno participó en el Programa 18-25 y 0 en caso contrario); $\widehat{D}_i$ representa el tratamiento estimado; I_i es el instrumento empleado para dicha estimación (es igual a 1 si el individuo tenía una edad igual o inferior a 25 años y 0 en caso contrario); X_i es la variable de asignación (edad del alumno a 31 de diciembre de 2013); y Z_i recoge las variables de control (como características socioeconómicas se incorporan los ingresos medios de la unidad familiar; el entorno del alumno se delimita mediante la tasa de desempleo del municipio donde el alumno asiste a clase y si vive en una localidad considerada rural; por último, se agrega información sobre el número de módulos en los que está matriculado).

Para llevar a cabo la evaluación hemos contado con información obtenida a partir de encuestas realizadas durante las primeras semanas del curso a los alumnos matriculados en educación de adultos. Simultáneamente, el Servicio de Enseñanza de Personas Adultas y a Distancia del Gobierno de Extremadura nos proporcionó información administrativa complementaria. Una vez fusionada la información disponible, restringimos la muestra para centrar el análisis en: i) la población masculina; ii) matriculada en al menos cuatro módulos; y iii) con edades comprendidas entre 20 y 31 años.

El cuadro n.º 1 recoge los resultados de los dos modelos estimados para estudiar el impacto del Programa 18-25. El Modelo 1 no tiene en cuenta las variables de control, mientras que el Modelo 2 sí las incluye para reducir los posibles sesgos derivados de la amplitud de la horquilla. Además, puesto que dicha amplitud (seis años a cada lado del punto de corte) puede dar lugar a que la muestra esté formada por personas con características muy diferentes (individuales, sociales y económicas), repetimos el análisis reduciendo la horquilla, en primer lugar, a cuatro años por encima y por debajo del punto de corte (todos aquellos con edades comprendidas entre los 22 y 29 años) y, en segundo lugar, a los varones entre 24 y 27 años (dos años a cada lado del punto de corte).

De acuerdo con el Modelo 1, el tratamiento resulta no ser estadísticamente significativo en ninguna de las horquillas de edad, observándose resultados similares en el Modelo 2, por lo que podemos concluir que, incluso corrigiendo por el posible sesgo existente mediante la incorporación de variables de control, el impacto del tratamiento continúa siendo no estadísticamente significativo. En definitiva, los resultados muestran que la probabilidad de obtener el título de la ESO no fue estadísticamente diferente para el grupo de tratados y para el grupo de control y, en consecuencia, el hecho de participar en el Programa 18-25 no supu-

CUADRO N.º 1

IMPACTO DEL PROGRAMA 18-25 SOBRE LA PROBABILIDAD DE TITULAR

HORQUILLA	*DE 24 A 27 AÑOS*		*DE 22 A 29 AÑOS*		*DE 20 A 31 AÑOS*	
VARIABLES	*MODELO 1*	*MODELO 2*	*MODELO 1*	*MODELO 2*	*MODELO 1*	*MODELO 2*
Tratamiento	0,6102	0,4830	-0,0306	-0,0442	-0,1100	-0,1285
	(0,4992)	(0,3945)	(0,2332)	(0,2232)	(0,1910)	(0,1913)
Edad	0,2556*	0,2424**	0,0222	0,0223	0,0051	0,0054
	(0,1329)	(0,1099)	(0,0318)	(0,0325)	(0,0186)	(0,0192)
Ingresos 1.000-1.500€		0,1841		-0,0177		0,0057
		(0,1528)		(0,1111)		(0,0865)
Ingresos >1.500€		0,4966***		0,0981		-0,0735
		(0,1864)		(0,1812)		(0,1357)
Tasa desempleo		0,0263***		0,0040		0,0063
		(0,0100)		(0,0074)		(0,0064)
Rural		-0,1042		0,0106		-0,0258
		(0,1085)		(0,0750)		(0,0617)
Módulos matriculados		-0,1768**		-0,1083*		-0,0687
		(0,0802)		(0,0615)		(0,0496)
Constante	-6,4226*	-5,7688*	-0,1761	0,3143	0,2851	0,5109
	(3,6319)	(2,8337)	(0,9049)	(0,9476)	(0,5493)	(0,5938)
Observaciones	100		190		279	

Nota: Errores estándar entre paréntesis. ***significativo al 1 %; **5 %, *10 %.
Fuente: Elaboración propia.

so, para los hombres comprendidos en las edades evaluadas, un incentivo suficiente.

V. CONCLUSIONES

El análisis realizado ha tratado de poner de relieve algunas de las principales aproximaciones que habitualmente se utilizan en el ámbito educativo para evaluar políticas o intervenciones dirigidas a mejorar los resultados académicos de los alumnos, como requisito previo para poder contar en el futuro con una población más y mejor formada que garantice mayores tasas de crecimiento y desarrollo económico.

El método ideal para realizar evaluaciones de políticas educativas es el desarrollo de experimentos aleatorios que permiten comprobar si la implementación de una medida ha tenido los efectos deseados mediante la comparación de resultados entre aquellos a los que se aplica la política educativa (grupo tratado) y a los que no (grupo de control). Su uso es cada vez más habitual, especialmente en algunos países anglosajones en los que su aplicación se ha convertido en requisito a la hora de implementar tales políticas. En los últimos años resulta más frecuente encontrar ejemplos de su empleo también en países en vías de desarrollo, llegando la labor de tres investigadores en este campo (Esther Duflo, Abhijit Banerjee y Michael Kremer) a ser reconocida en 2019 con un Premio Nobel por su labor de fomento del uso de ensayos aleatorios en aquellos países. En España, sin embargo, la utilización de este tipo de diseños es todavía una asignatura pendiente, aunque hay que reconocer su promoción por parte de algunas instituciones privadas que han puesto en marcha algunas pruebas piloto descritas en este artículo. En todo caso, es preciso una mayor implicación de las instituciones públicas en la necesidad de evaluar las intervenciones con el fin de mejorar su diseño. El papel que en la actualidad está desarrollando la Autoridad Independiente de Responsabilidad Fiscal (AIReF) con la creación este año de la división de evaluación del gasto público va en la buena dirección.

Cuando no es posible utilizar experimentos aleatorios por razones diversas, fundamentalmente éticas y económicas, los investigadores recurren al uso de técnicas cuasiexperimentales con las que tratan de reproducir las condiciones ideales de los experimentos aleatorios corrigiendo

los habituales problemas de endogeneidad. Su correcta implementación se facilita si se dispone de datos administrativos o censales (Connelly *et al.,* 2016) con los que poder analizar a toda la población de interés aprovechando las variaciones existentes entre los diferentes grupos que componen la base de datos, o para crear cohortes de individuos con el fin de estudiar los cambios que se producen entre dos períodos y reunir información sobre individuos que han vivido un determinado cambio (repetición de curso, modificación en la regulación del sistema educativo en diferentes zonas geográficas, etc.).

En España, la responsabilidad de recopilar y codificar este tipo de datos sobre los alumnos de educación primaria y secundaria de manera periódica recae sobre las comunidades autónomas, las responsables de su prestación. Su explotación en estudios empíricos, con técnicas de inferencia causal, es todavía muy escasa ya que no están disponibles, con generalidad, para la comunidad académica por los impedimentos que ponen tales administraciones educativas.

Una alternativa a los datos de carácter censal en la evaluación de políticas educativas son los datos muestrales proporcionados por diferentes evaluaciones nacionales e internacionales de conocimientos en distintas materias a gran escala, una práctica que se ha vuelto cada vez más habitual tanto en los estudios desarrollados en el ámbito internacional como especialmente en el caso de España, dada la ausencia o las dificultades de acceso a mejores fuentes de información. A pesar de que estos datos no fueron diseñados originalmente para la implementación de técnicas de inferencia causal, en la literatura pueden encontrarse múltiples ejemplos en los que los investigadores tratan de estimar efectos causales de programas o intervenciones educativas a partir de ese tipo de datos (Cordero *et al.,* 2018).

Por último, en relación con el ejercicio de evaluación presentado, el Programa 18-25 implementado por el Gobierno de Extremadura a principios de la década anterior con el propósito de mejorar el nivel de formación de los desempleados mediante un incentivo económico, los resultados obtenidos aplicando la técnica de regresión en discontinuidad muestran que la probabilidad de obtener el título de la ESO no fue estadísticamente diferente para el grupo de individuos que participaron en el programa respecto a los que no participaron. Por tanto, se puede concluir que el Programa 18-25 no resultó efectivo. Este resultado es similar al obtenido en otra investigación realizada por los mismos autores con un diseño similar, aunque con alguna variación en los datos empleados (Pedraja *et al.,* 2022).

NOTAS

(1) Con resultados del informe PISA, aproximadamente tres cuartas partes de la variación de las tasas de crecimiento entre países pueden explicarse por dos variables: el nivel inicial de ingresos y los resultados académicos de la población.

(2) Este artículo resume los principales resultados obtenidos en más de 100 estudios sobre la función de producción educativa realizados en Estados Unidos.

(3) Los interesados en los problemas que origina la endogeneidad en la estimación de modelos econométricos en el ámbito educativo y en los principales métodos experimentales y cuasiexperimentales para la identificación de efectos causales, pueden consultar los trabajos de Webbink (2005) o Schlotter *et al.* (2011).

(4) Una detallada explicación de estas técnicas se encuentra en el trabajo de Artés y Rodríguez-Sánchez (2022) en este mismo número o en alguno de los manuales sobre evaluación de impacto existentes en la literatura, entre otros, Angrist y Pischke (2014), Khandker *et al.* (2010) o Gertler *et al.* (2016).

(5) No incluimos dentro de esta categoría a los modelos de *matching* o emparejamiento (entre los que se incluye el *propensity score matching*) porque, aunque reducen los sesgos que genera el problema de la autoselección, no corrigen los sesgos asociados a la existencia de diferencias no observadas entre el grupo tratado y el de control.

(6) En Estados Unidos existe una Evaluación Nacional del Progreso Educativo (NAEP, por sus siglas en inglés) desde 1969. Además, desde mediados de los años noventa, muchos Estados disponen de sistemas de evaluación periódica de los conocimientos de los estudiantes que permiten un seguimiento longitudinal y emparejarlos con sus escuelas y, en algunos casos, con sus profesores. Los tres Estados que cuentan con los sistemas más antiguos y que han estado ampliamente disponibles para los investigadores son Florida, Carolina del Norte y Texas. En Europa, los países nórdicos cuentan con registros administrativos desde mediados de los años sesenta, en Reino Unido existe una base de datos censal (NPD) desde mediados de los noventa, y en Italia (INVALSI) desde 2005.

(7) Sin embargo, la mayoría de los programas de incentivos al profesorado que se han instaurado en países en vías de desarrollo, donde el margen de mejora es mucho más amplio y los salarios de los docentes son más bajos, han tenido un impacto muy positivo, consiguiendo generar mejoras importantes en los resultados de los estudiantes (Ganimian y Murname, 2016).

(8) Véase Mislevy *et al.* (1992) o Wu (2005) para una explicación detallada del concepto de valores plausibles y los procedimientos empleados para su construcción.

(9) Los trabajos de Van der Klaauw (2008) y Cattaneo *et al.* (2018) ofrecen explicaciones detalladas de este enfoque metodológico y muchos ejemplos prácticos sobre la aplicación de la regresión en discontinuidad.

(10) Este trabajo es una actualización de otro muy influyente en la literatura publicado veinte años antes (Angrist y Lavy, 1999), en el que, utilizando una estrategia de identificación muy parecida, los autores sí encontraron un efecto positivo de la reducción del tamaño de clase sobre los resultados escolares.

(11) Los trabajos de Murnane y Willet (2011) y Pokropek (2016) ofrecen una excelente discusión acerca de esta cuestión y clasifican las fuentes más populares de posibles instrumentos de investigación en el ámbito educativo.

(12) Para una revisión de los determinantes del rendimiento académico a partir de los datos de PISA, véase Cordero *et al.* (2013).

(13) Para más información, véase: https://educaixa.org/es/evaluacion-programas-educativos

BIBLIOGRAFÍA

Anghel, B., Cabrales, A. y Carro, J. M. (2016). Evaluating a bilingual education program in Spain: The impact beyond foreign language learning. *Economic Inquiry*, 54(2), pp. 1202-1223.

Angrist, J., Bettinger, E. y Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review*, 96(3), pp. 847-862.

Angrist, J. D. y Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2), pp. 533-575.

Angrist, J. D., Lavy, V., Leder-Luis, J. y Shany, A. (2019). Maimonides' rule redux. *American Economic Review: Insights*, 1(3), pp. 309-324.

Angrist, J. D. y Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4), pp. 979-1014.

Angrist, J. D. y Pischke, J. S. (2014). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Arenas, A. y Gortazar, L. (2022). *Learning loss One Year After School Closures: Evidence from the Basque Country*. Working Paper #1, Esade Center for Economic Policy.

Artés, J. y Rodríguez-Sánchez, B. (2022). Métodos de evaluación de políticas públicas. *Papeles de Economía Española*, 172, pp. 18-29..

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics*, 30, pp. 143-153.

Bover, O., Hospido, L. y Villanueva, E. (2018). The impact of high school financial education on financial knowledge and choices: Evidence from a randomized trial in Spain. *IZA Discussion Papers, n.º 11265.* Institute of Labor Economics (IZA), Bonn.

Briole, S., Gurgand, M., Maurin, É., McNally, S., Ruiz-Valenzuela, J. y Santín, D. (2022). The Making of Civic Virtues: a school-based experiment in three countries. *IZA Discussion Papers,* n.º 15141. Institute of Labor Economics (IZA), Bonn.

Cabras, S. y Tena, J. D. (2016). A Bayesian non-parametric modeling to estimate student response to ICT investment. *Journal of Applied Statistics*, 43(14), pp. 2627-2642.

Cattaneo, M. D., Idrobo, N. y Titiunik, R. (2018). *A Practical Introduction to Regression Discontinuity Designs, Vol. I y II.* Cambridge: Cambridge University Press.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. y Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126(4), pp. 1593-1660.

Choi, Á., Gil, M., Mediavilla, M. y Valbuena, J. (2018). The evolution of educational inequalities in Spain: dynamic evidence from repeated cross-sections. *Social Indicators Research*, 138(3), pp. 853-872.

Coleman, J., Campbell, E., Hobson, C., Mc Partland, I., Modd, A., Wenfeld, F. y York, R. (1966). *Equality of educational opportunity*. Department of Health, Education and Welfare, Office of Education, Government Printing Office, Washington.

Connelly, R., Playford, C. J., Gayle, V. y Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, pp. 1-12.

Cordero, J. M., Crespo-Cebada, E. y Pedraja Chaparro, F. (2013). Rendimiento educativo y determinantes según PISA: Una revisión de la literatura en España. *Revista de Educación*, 362, pp. 273-297.

Cordero, J. M., Cristóbal, V. y Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, 32(3), pp. 878-915.

Cordero, J. M. y Gil-Izquierdo, M. (2018). The effect of teaching strategies on student achievement: An analysis using TALIS-PISA-link. *Journal of Policy Modeling*, 40(6), pp. 1313-1331.

Creemers, B. y Kyriakides, L. (2007). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.

Ding, Y., Lu, F. y Ye, X. (2020). Intergovernmental transfer under heterogeneous accountabilities: the effects of the 2006 Chinese education finance reform. *Economics of Education Review*, 77, 101985.

Duflo, E. y Kremer, M. (2005). Use of Randomization in the Evaluation of Development Effectiveness. En O. Feinstein, G. K. Ingram y G. K. Pitman (eds.), *Evaluating development effectiveness*, pp. 205-232. New Brunswick, New Jersey and London: Transaction Publishers.

Eberts, R., Hollenbeck, K. y Stone, J. (2002). Teacher performance incentives and student outcomes. *Journal of Human Resources*, 37(4), pp. 913-927.

Felfe, C., Nollenberger, N. y Rodríguez-Planas, N. (2015). Can't buy mommy's love? universal childcare and children's long-term cognitive development. *Journal of Population Economics*, 28(2), pp. 393-422.

Finn, J. D. y Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), pp. 97-109.

Ganimian, A. J. y Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact

evaluations. *Review of Educational Research*, 86(3), pp. 719-755.

García-Pérez, J. I., Hidalgo-Hidalgo, M. y Robles-Zurita, J. A. (2014). Does grade retention affect students' achievement? Some evidence from Spain. *Applied Economics*, 46, pp. 1373-1392.

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B. y Vermeersch, C. M. (2016). *Impact evaluation in practice.* The World Bank.

Glewwe, P., Kremer, M. y Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied economics*, 1(1), pp. 112-135.

Greene, J. P. y Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, 2(4), pp. 319-340.

Gustafsson, J. E. (2013). Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, 24(3), pp. 275-295.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3), pp. 1141-1177.

Hanushek, E. A. y Woessmann, L. (2011). The economics of international differences in educational achievement. En E. Hanushek, S. Machin y L. Woessmann (eds.), *Handbook of the Economics of Education, 3,* pp. 89-200. Amsterdam: North Holland.

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

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. y Yavitz, A. (2010b). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1(1), pp. 1-46.

Hidalgo, A. y Lopez-Mayan, C. (2018). Teaching styles and achievement: Student and teacher perspectives. *Economics of Education Review*, 67, pp. 184-206.

Imberman, S. A. y Lovenheim, M. F. (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics*, 97(2), pp. 364-386.

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

Jerrim, J., López-Agudo, L. A., Marcenaro-Gutiérrez, O. D. y Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, 61, pp. 51-58.

Jerrim, J., López-Agudo, L. A. y Marcenaro-Gutiérrez, O. D. (2019). The relationship between homework and the academic progress of children in Spain during compulsory elementary education: A twin fixed-effects approach. *British Educational Research Journal*, 45(5), pp. 1021-1049.

Jerrim, J., López-Agudo, L. A. y Marcenaro-Gutiérrez, O. D. (2020). Does it matter what children read? New evidence using longitudinal census data from Spain. *Oxford Review of Education*, 46(5), pp. 515-533.

Jürges, H., Schneider, K. y Büchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association*, 3(5), pp. 1134-1155.

Khandker, S. R., Koolwal, G. B. y Samad, H. A. (2010). *Handbook on impact evaluation: quantitative methods and practices.* Washington DC: World Bank.

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review*, 93(2), pp. 102-106.

Laajaj, R., Moya, A. y Sánchez, F. (2022). Equality of opportunity and human capital accumulation: Motivational effect of a nationwide scholarship in Colombia. *Journal of Development Economics*, 154, 102754.

Lai, F., Sadoulet, E. y De Janvry, A. (2011). The contributions of school quality and teacher qualifications to student performance evidence from a natural experiment in Beijing middle schools. *Journal of Human Resources*, 46(1), pp. 123-153.

Lee, J. y Fish, R. M. (2010). International and interstate gaps in value-added math achievement: Multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education*, 117(1), pp. 109-137.

López-Agudo, L. A. y Marcernaro-Gutiérrez, O. D. (2019a). The Effect of Weekly Instruction Time on Academic Achievement: The Spanish Case. *Hacienda Pública Española*, 230, pp. 63-93.

López-Agudo, L. A. y Marcenaro-Gutiérrez, O. D. (2019b). Are Spanish children taking advantage of their weekly classroom time? *Child Indicators Research*, 12(1), pp. 187-211.

López-Agudo, L. A. y Marcenaro-Gutiérrez, O. D. (2021). La falta de influencia del tamaño de la clase sobre el rendimiento académico de los estudiantes: evidencia empírica para Andalucía. *Revista de Educación*, 395, pp. 321-361.

López-Rupérez, F., García-García, I. y Expósito, E. (2021). La repetición de curso y la graduación en Educación Secundaria Obligatoria en España: análisis empíricos y recomendaciones políticas. *Revista de Educación*, 394, pp. 325-353.

López-Torres, L., Prior, D. y Santín, D. (2019). Assessing the effect of educational programs on public schools' performance. *Applied Economics*, 51(48), pp. 5205-5226.

LUCAS, A. M. y MBITI, I. M. (2012). Access, sorting, and achievement: The short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, 4(4), pp. 226-253.

MACHIN, S. y MCNALLY, S. (2008). The literacy hour. *Journal of Public Economics*, 92(5-6), pp. 1441-1462.

MACHIN, S., MCNALLY, S. y SILVA, O. (2007). New technology in schools: Is there a payoff? *The Economic Journal*, 117(522), pp. 1145-1167.

MANCEBÓN, M. J., PÉREZ-XIMÉNEZ DE EMBÚN, D. y VILLAR-ALDONZA, A. (2018). Evaluación del efecto de la escolarización temprana sobre las habilidades cognitivas y no cognitivas de los niños de cinco/seis años. *Hacienda Pública Española*, 226(3), pp. 123-153.

MEDIAVILLA, M., MANCEBÓN, M. J., GÓMEZ-SANCHO, J. M. y JIMÉNEZ, L. P. (2019). Bilingual education and school choice: A case study of public secondary schools in the Spanish region of Madrid. *IEB Working Paper,* n.º 2019/01 Institut d'Economia de Barcelona.

MISLEVY, R. J., BEATON, A. E., KAPLAN, B. y SHEEHAN, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), pp. 133-161.

MURNANE, R. J. y WILLETT, J. B. (2011). *Methods matter*. Oxford University Press. New York.

PEDRAJA, F., SANTÍN, D. y SIMANCAS, R. (2016). The impact of immigrant concentration in schools on grade retention in Spain: a difference-in-differences approach. *Applied Economics*, 48(21), pp. 1978-1990.

PEDRAJA, F., SANTÍN, D. y SIMANCAS, R. (2022). Show me the money! The impact of a conditional cash transfer on educational achievement. *Empirical Economics*, en prensa. doi: https://doi.org/10.1007/s00181-022-02211-x

PFEFFERMANN, D. y LANDSMAN, V. (2011) Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied* Statistics, 5(3), pp. 1726-1751.

POKROPEK, A. (2016). Introduction to instrumental variables and their application to large-scale assessment data. *Large-scale Assessments in Education*, 4(1), pp. 1-20.

SALINAS, J. y SANTÍN, D. (2012). Selección escolar y efectos de la inmigración sobre los resultados académicos españoles en PISA 2006. *Revista de Educación*, 358, pp. 382-405.

SANTÍN, D. y SICILIA, G. (2015). El impacto de la educación infantil en los resultados de primaria: evidencia para España a partir de un experimento natural. En: D. SANTÍN PAU BALART, A. CABRALES, J. CALERO, Á. DE LA FUENTE, J. ORIOL ESCARDÍBUL, F. FELGUEROSO, G. SICILIA (eds.), *Reflexiones sobre el Sistema Educativo Español*, pp. 45 -74. Madrid: Editorial Centro de Estudios Ramón Areces S. A.

SCHLOTTER, M., SCHWERDT, G. y WOESSMANN, L. (2011). Econometric methods for causal evaluation of education policies and practices: a non-technical guide. *Education Economics*, 19(2), pp. 109-137.

SCHWERDT, G. y WUPPERMANN, A. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review*, 30, pp. 365-379.

TODD, P. E. y WOLPIN, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), pp. 3-33.

THISTLEWAITE, D. y CAMPBELL, D. (1960). Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51, pp. 309-317.

VAN DER KLAAUW, W. (2008). Regression-discontinuity analysis: a survey of recent developments in economics. *Labour*, 22(2), pp. 219-245.

VANDENBERGHE, V. y ROBIN, S. (2004). Evaluating the effectiveness of private education across countries: a comparison of methods. *Labour Economics*, 11(4), pp. 487-506.

VILLAPLANA, C. (2014). Evaluación del programa Escuela 2.0 para la asignatura de Matemáticas a partir de PISA 2012. *Investigaciones de Economía de la Educación*, 9, pp. 631-652.

WEBBINK, D. (2005). Causal effects in education. *Journal of Economic Surveys*, 19(4), pp. 535-560.

WORD, E., JOHNSTON, J., BAIN, H. P., FULTON, B., ZAHARIES, J. B., LINTZ, M. N., ACHILLES, C. M., FOLGER, J. y BREDA, C. (1990). *Student/ Teacher Achievement Ratio (STAR), Tennessee's K-3 Class Size Study: Final Summary Report*, 1985-1990. Tennessee State Department of Education, Nashville.

WU, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2), pp. 114-128.