RMIE, 2019, VOL. 24, NÚM. 82, PP 871-878 (ISSN: 14056666 • ISSN-e 25942271)

Aporte de discusión

LA INCAPACIDAD DEL BIG DATA DE ESCAPAR DE LAS LIMITACIONES DE LA EVALUACIÓN DE IMPACTO*

BRENT EDWARDS JR

Resumen:
Este artículo analiza los métodos que se utilizan para llevar a cabo "evaluaciones de impacto", un tipo de investigación que supuestamente puede determinar el efecto de una política o programa en los resultados educativos. Gracias al aumento de la recopilación de datos sobre estudiantes y escuelas en el campo de la educación, estos métodos se han vuelto cada vez más populares, a pesar de sus serias limitaciones. Este trabajo aborda las deficiencias de los dos enfoques de evaluación de impacto más comunes: el análisis de regresión y los experimentos aleatorios (*Randomized Controlled Trials*, RCT). Se advierte que, aunque las grandes cantidades de datos y los métodos de evaluación de impacto pueden ser insumos útiles para las discusiones sobre políticas, es importante que sus límites sean comprendidos.

Abstract:
This article analyzes the methods used to carry out "impact evaluations", a type of research that supposedly can determine the effect of a policy or program on educational results. Thanks to the increased compilation of data on students and schools in the field of education, these methods have become ever more popular, in spite of their serious limitations. The study addresses the deficiencies of the two most common focuses of impact evaluations: regression analysis and random experiments (Randomized Controlled Trials, or RCT). The observation is that although large amounts of data and the methods of impact evaluation can be useful inputs for policy discussions, it is important to understand their limits.

Palabras clave: investigación educativa; método cuantitativo; análisis de datos; toma de decisiones; políticas públicas.
Keywords: educational research; quantitative method; data analysis; decision making; public policies.

Brent Edwards Jr: profesor en la University of Hawai´i at Mānoa, Honolulu, Hawái, Estados Unidos. CE: dbrente@gmail.com

*Este texto se basa en el libro *Global education policy, impact evaluations, and alternatives: The political economy of knowledge production*, de Brent Edwards Jr., publicado en 2018.

Introducción

Vivimos en la era de los macrodatos (*Big Data*), donde los sistemas escolares de todo el mundo se están "ahogando" en información, gracias a iniciativas como el Programa Internacional de Evaluación de Alumnos (PISA) y otras evaluaciones de rendimiento académico que reúnen una cantidad increíble de datos de puntajes de exámenes, estudiantes, familias y escuelas (Gorur; Sellar y Steiner-Khamsi, 2018). Sin embargo, no se ha prestado suficiente atención a los métodos cuantitativos que se utilizan para procesar y trasformar estos datos, para llegar a hallazgos relacionados con "lo que funciona". Aunque la idea del *Big Data* y su procesamiento están recibiendo cada vez más atención, el punto subyacente es que estas nuevas iniciativas y avances en la recopilación de datos aún dependen de métodos que tienen limitaciones importantes. Estos métodos se incluyen bajo la etiqueta de "evaluaciones de impacto" y pueden agruparse en dos categorías: análisis de regresión y experimentos aleatorios (*randomized controlled trials*; en adelante RCT); este último se considera el patrón de oro para revelar los vínculos causales entre las intervenciones y sus resultados.

Los problemas de los análisis de regresión

La idea básica del análisis de regresión es que un investigador pueda identificar el impacto de un determinado programa o variable en los resultados de interés (por ejemplo, los puntajes de las pruebas de los estudiantes). Para hacerlo, sin embargo, el académico debe controlar todas las variables que pudieran afectar el resultado. Aun así, en realidad rara vez, si es que alguna vez, es posible incluir todas las variables relevantes como controles estadísticos. Además, las variables incluidas deben medirse de manera apropiada y sus interrelaciones deben modelarse correctamente (Klees, 2016). A pesar de la incapacidad de cumplir con estos requisitos, ciertos estudiosos continúan empleando los análisis de regresión. Y no es que los investigadores no estén conscientes de estas limitaciones. Tomemos, por ejemplo, el documento de un economista senior de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) sobre el impacto del uso de la computadora en los puntajes de PISA. El autor escribe: "la comparación entre los que usan la computadora [...] y los que no [...] sería legítima solo si los estudiantes en estos dos grupos tuvieran características similares.

Desafortunadamente, hay [...] razones de peso para suponer que esto no es el caso" (Spiezia, s/a). Luego, el autor explica que "una correlación positiva entre el uso de la computadora y el rendimiento del estudiante puede ser que simplemente capture los efectos de un mejor entorno familiar" (p. 1), con la implicación de que esto es un problema porque no es posible controlar todas las formas en que "un mejor entorno familiar" afecta, en principio, el uso de la computadora y, posteriormente, las puntuaciones de PISA. O, en palabras del autor, "las diferencias observadas en el uso de la computadora ... pueden reflejar diferencias no observables en las características de los estudiantes" (p. 2). A pesar de ello, después de reconocer estos puntos en la introducción, el autor continúa con el análisis y ofrece implicaciones de política.

Las limitaciones de los RCT

En cuanto a los RCT, la idea es que uno puede determinar el efecto de una intervención asignando al azar a los participantes a los grupos de control y tratamiento, y luego observando la diferencia en los resultados promedio de cada grupo. Si se realiza correctamente, la diferencia de resultados entre los grupos es atribuible a la intervención, ya que la selección aleatoria debe hacer que los grupos de control y de tratamiento sean iguales en términos de todas las características relevantes. Sin embargo, en la práctica rara vez se logra una muestra equilibrada y representativa de la población de interés a través de la selección aleatoria, ya que siempre hay diferencias en una u otra variable. Por ejemplo, cuando las comunidades se asignan al azar a grupos de tratamiento y control, pueden ser equivalentes en términos de algunos antecedentes, pero hay muchas variables —relacionadas con los estudiantes, sus familias, sus comunidades y los contextos en los que están integrados— que afectan los resultados. Es necesario que los dos grupos sean idénticos en todos los aspectos, porque cuando no lo son, ya no pueden atribuirse los resultados observados a la intervención que se estudia, que es el punto central de los RCT.

Otro problema tiene que ver con la imparcialidad de las estimaciones. Si asumimos que la selección al azar ha funcionado, entonces "la diferencia entre las medias de los tratamientos y los controles es una estimación del efecto promedio del tratamiento entre los participantes", como ha señalado el economista de Princeton, Angus Deaton (2010). Sin embargo,

aunque este hecho es a menudo la base del interés en los RCT, lo que este enfoque no resalta es que la selección aleatoria no necesariamente significa que los coeficientes de las variables incluidas sean imparciales en cada experimento, solo que lo son, en promedio, en todos los experimentos replicados. Como Angus Deaton ha señalado también: "La imparcialidad dice que, si repitiéramos el experimento muchas veces, tendríamos razón en promedio. Sin embargo, casi nunca estamos en una situación así, y con un solo experimento (como casi siempre es el caso), la imparcialidad no previene que nuestra única estimación esté muy lejos de la verdad" (Deaton y Cartwrigh, 2016).

Una tercera debilidad importante de los RCT es que solo proporcionan estimaciones de las medias de los efectos del tratamiento, dado que los resultados de los individuos en los grupos de intervención y control se promedian. La diferencia entre las medias se toma entonces como el efecto promedio (esperando sea en beneficio) para quienes recibieron el tratamiento. Sin embargo, este promedio podría estar encubriendo una situación en la que "casi toda la población está herida y unos pocos reciben beneficios muy grandes" (Deaton, 2010:439). Esta situación claramente presenta problemas para quienes toman decisiones, ya sea para un médico que debe decidir qué es lo mejor para un paciente o para un político que debe decidir sobre política pública. La presencia de un valor atípico en el grupo de tratamiento o en el de control puede sesgar los promedios generales y puede llevar a la implementación (o no) de un programa que podría ser beneficioso para todos (o que podría hacer daño).

Cuarto, y finalmente, hay problemas cuando se trata de la generalización de los resultados derivados del RCT; si bien estos se esfuerzan por garantizar que sus resultados tengan validez interna, no es posible decir algo automáticamente acerca de la transferencia de los resultados a otros contextos. Resulta interesante, e irónico, que defender la aplicabilidad de los resultados en otros contextos requiera que los académicos recurran a los tipos de comparaciones y métodos cualitativos que los defensores de los RCT intentan evitar. Estos están diseñados para identificar el efecto causal de las intervenciones para muestras particulares; su diseño no garantiza que los resultados sean transferibles. Cuando se trata de la transferencia de políticas, los RCT no son más útiles que otras formas de investigación, incluso puede que lo sean menos, ya que no son generalizables y los resultados no dicen nada acerca la implementación del programa.

Se sugiere que tanto los productores como los consumidores de RCT se esfuercen para obtener una comprensión crítica del contexto de la evaluación, es decir, una comprensión que vaya más allá de un nivel superficial, para incluir aspectos socioculturales, estructurales, históricos y políticos. Una comprensión de esta naturaleza se considera necesaria para percibir las maneras en que estos aspectos del contexto (a menudo ignorados) se hacen presentes y pueden afectar, no solo el comportamiento de los participantes en el estudio, sino también los resultados del enfoque.

El Big Data depende de los métodos de evaluación de impacto

Sin duda, la tendencia del *Big Data* en el campo de la política educativa global confronta lo presentado anteriormente. Algunos ejemplos no solo aclaran este punto, sino que también enfatizan la conexión entre la evaluación de impacto y los macrodatos. Por ejemplo, el Banco Mundial ha ampliado su oferta de asesoramiento sobre políticas basado en los puntajes de las pruebas. Esto es, después de décadas de promover las pruebas estandarizadas, el Banco Mundial hizo del "aprendizaje para todos" la idea central de su documento de estrategia del sector educativo más reciente. Luego, dio seguimiento a esta idea mediante la creación de "la base de datos en panel comparable sobre calidad educativa más grande del mundo", que abarca los años 1965-2015 e incluye 163 países (World Bank, 2018). Y ahora, la OCDE y el Banco Mundial están trabajando juntos para ampliar el alcance de PISA, adaptándolo a países en "desarrollo" (es decir, países de ingresos medios y bajos). Esta nueva iniciativa, conocida como PISA para el Desarrollo (o PISA-D), permitirá que la educación sea comparada y regida por un solo examen, tanto en países pobres como ricos. Pero PISA-D es solo una de las iniciativas de la OCDE para expandir su influencia a través del "gobierno por números" (Grek, 2009). Con ese fin, también ha desarrollado pruebas para escuelas individuales (a diferencia de su enfoque tradicional en muestras a nivel de país), para medir competencias en adultos y en educación superior.

Si bien los datos de los ejemplos anteriores requieren manipulación por análisis de regresión para generar hallazgos sobre "lo que funciona", hay otras iniciativas que también incorporan RCT. Por ejemplo, la de la Universidad de Stanford que busca hacer que la guerra contra la pobreza sea "inteligente", ha reunido un equipo multidisciplinario de profesores que "usarán el aprendizaje automático [*machine learning*] para seleccionar

información de entre grandes conjuntos de datos para comprender las muchas variables que conducen, perpetúan, y potencialmente incluso evitan el empobrecimiento" (Walsh, 2019). Como se describe en un artículo reciente sobre este proyecto, se busca reunir a científicos informáticos para "aplicar el análisis de datos y el aprendizaje automático a la vasta red de información recopilada", con el objetivo final de aplicar "modelos predictivos que sugieran qué intervenciones tienen probabilidad de funcionar mejor en un contexto determinado" (Walsh, 2019). Metodológicamente, para seleccionar los conjuntos de datos e identificar variables asociadas con la pobreza, indudablemente usarán algún tipo de análisis de regresión. Esta fase será seguida por otra en la que el equipo diseñe y pruebe, a través de un RCT, las intervenciones que podrían ayudar a las personas a escapar de la pobreza.

Más allá de los métodos de evaluación de impacto

Si bien la idea del *Big Data* y la capacidad de su procesamiento están recibiendo cada vez más atención, se enfatiza el punto introductorio acerca de que estas nuevas iniciativas y avances en la recopilación de datos aún dependen de métodos que tienen limitaciones importantes.

Muchos proyectos lanzados por organizaciones internacionales junto con los gobiernos son ciertamente ambiciosos e impresionantes en su alcance, pero se presentan con una apariencia de certidumbre y objetividad que no merecen. Con ese fin, los defensores del *Big Data* evitan o minimizan la discusión de las dificultades metodológicas de la evaluación de impacto, ni tampoco reconocen la dinámica política y organizacional que afecta la recolección de datos. Por ejemplo, cuando los políticos asignan a las comunidades que participan en las intervenciones (en lugar de ser asignadas al azar), así como la interpretación de los hallazgos, que necesariamente está limitada por agendas institucionales y por la estructura de incentivos en la que está inmerso el investigador (Edwards, 2018).

Sin duda, los macrodatos y los métodos de evaluación de impacto pueden ser insumos útiles para las discusiones sobre políticas en el contexto de las reformas de políticas y la gobernanza global. Sin embargo, es esencial que sus limitaciones sean comprendidas por quienes los producen y los utilizan para las reformas de políticas, así como por los actores en todos los niveles. En la medida en que dichos métodos se utilicen cada

vez más para guiar la política pública en todo el mundo, es esencial que los actores interesados, dentro y fuera de los sistemas educativos, estén al tanto de sus debilidades metodológicas y de su incapacidad para eliminar la política de la formulación de políticas. En efecto, si bien las promesas del *Big Data* son seductoras, en general, no ha reemplazado el elemento humano de la toma de decisiones. Es decir, aunque los datos pueden construir una cierta visión del mundo, y la interpretación de esos datos, ya sea por un investigador o por un programa de computadora, puede sugerir ciertas medidas de política, es poco probable que las decisiones de políticas importantes se automaticen pronto. La implicación es, entonces, que los políticos y altos funcionarios y los métodos en los que se basan deben sujetarse a un estándar más alto, uno que vaya más allá de la combinación de *Big Data* y la evaluación de impacto. En lugar de llevar a la mejor o a las mejores prácticas, la intersección de los macrodatos y las técnicas de evaluación de impacto puede conducir a estudios y, posteriormente, a reformas que son costosas, perjudiciales, contextualmente irrelevantes y/o inefectivas.

La atención debe dirigirse a moderar el uso de estos métodos, complementándolos con otras estrategias (cualitativas) y cambiando la naturaleza de la toma de decisiones, de un ejercicio que busca ser tecnocrático a uno que sea abierta e inevitablemente político. El *Big Data* no genera políticas sin política, más bien, oculta gran parte de los sesgos y el aspecto político del proceso con una apariencia de tecnificación. En el contexto actual de la gobernanza global, que se encuentra bajo la influencia de instituciones internacionales y los gobiernos nacionales que cooperan con ellas, es probable que esta situación no produzca beneficios significativos para los más marginados, ya que ninguno de estos actores se moviliza por un cálculo político que responda a las necesidades de los pobres. Comprender de manera crítica las limitaciones del *Big Data* y de las evaluaciones de impacto es solo un primer paso para desafiar el estatus quo de la gobernanza global en general, es decir, pensar más allá de los métodos, la política y las cosmovisiones que actualmente impregnan y limitan el campo de la política educativa global.

Traducción: Este texto fue traducido del inglés por Dulce Lomelí.

Referencias

Deaton, Angus (2010). "Instruments, randomization, and learning about development", *Journal of Economic Literature,* vol. 48, núm. 2, pp. 424-455. Disponible en: https://www.princeton.edu/~deaton/downloads/deaton%20instruments%20randomization%20learning%20about%20development%20jel%202010.pdf

Deaton, Aangus y Cartwright, Nancy (2016). "The limitations of randomized controlled trials", *Vox*, 9 de noviembre. Disponible en: http://voxeu.org/article/limitations-randomised-controlled-trials

Edwards Jr., D. Brent (2018). *Global education policy, impact evaluations, and alternatives: The political economy of knowledge production*, Nueva York: Palgrave MacMillan. Disponible en: https://www.academia.edu/36174726/Global_Education_Policy_Impact_Evaluations_and_Alternatives_The_Political_Economy_of_Knowledge_Production?source=swp_share

Gorur, Radhika; Sellar, Sam y Steiner-Khamsi, Gita (2018). "Big Data and even bigger consequences", en R. Gorur, S. Sellar y G. Steiner-Khamsi (eds.), *World Yearbook of Education 2019: Comparative methodology in the era of Big Data and global networks*, Nueva York: Routledge.

Grek, Sotiria (2009). "Governing by numbers: The PISA 'effect' in Europe", *Journal of Education Policy*, vol. 24, núm. 1, pp. 23-37. Disponible en: https://www.research.ed.ac.uk/portal/files/14608705/Governing_by_Numbers.pdf

Klees, Steven (2016). "Inferences from regression analysis: Are they valid?", *Real-world Economics Review*, vol. 74, pp. 85-97. Disponible en: http://www.paecon.net/PAEReview/issue74/Klees74.pdf

Spiezia, Vincenzo (s/d). *Does computer use increase educational achievements? Student-level evidence from PISA*. Disponible en: https://pdfs.semanticscholar.org/f388/69cbf15b240aae015923391c6cecbe5eacc2.pdf

Walsh, Dylan (2019). "Solving poverty using the tools of Silicon Valley", *Stanford Graduate School of Business* (sitio web), 9 de enero. Disponible en: https://stanford.io/2Ri9mdP

World Bank (2018). *Global data set on education quality*, Washington, DC. Disponible en: http://datatopics.worldbank.org/education/wQueries/qachievement

Recibido: 10 de mayo de 2019
Aceptado: 15 de julio de 2019