illustro | REVISTA DE INVESTIGACIÓN EN CIENCIAS ECONÓMICAS, CONTABLES Y EMPRESARIALES

Revisión sistemática de literatura teórica sobre métodos de evaluación de impacto para evaluar políticas públicas

Systematic review of theoretical literature on impact evaluation methods to evaluate public policies

Christian Rivadeneira Vildoso
Pontificia Universidad Católica del Perú • christian.rivadeneira@pucp.edu.pe

Resumen

Los programas y las políticas públicas se diseñan e implementan con el objetivo de cambiar algún comportamiento o aspecto de la vida de una persona, como su educación, sus ingresos, su alimentación, etc. Esta implementación se basa en la teoría del cambio para lograr sus objetivos. Sin embargo, determinar si estos programas o políticas públicas logran los impactos o resultados esperados es una labor que a menudo no es abordada por los implementadores de las políticas públicas. La evaluación de impacto de las políticas públicas es una acción fundamental dentro de la gestión pública, pues permite obtener información sobre los impactos que tienen dichas políticas o programas en la población beneficiaria y así poder estudiar la posibilidad de ampliar el alcance, modificar, sugerir cambios o detener la ejecución. En esta revisión sistemática se evaluaron 22 estudios sobre evaluación de impacto. La temporalidad de los estudios evaluados va desde el año 1983 hasta el año 2017. Uno de los principales hallazgos está referido a la importancia de estimar un contrafactual creíble que permita identificar de manera adecuada el impacto de la aplicación de una determinada política pública. Otro de los aspectos a resaltar es que la metodología para evaluar una determinada política pública depende de los datos con los que se cuenta y de las características de la intervención. Finalmente, los modelos econométricos utilizados para estimar los impactos no son, en ningún caso, modelos rígidos. Mucho dependerá de la habilidad y la creatividad del investigador para plantear una propuesta metodológica que cumpla con el objetivo planificado.

Palabras clave: evaluación de impacto, sesgo de selección, emparejamiento, diferencias en diferencias, discontinuidad, efecto del tratamiento.

https://doi.org/10.36901/illustro.v11i.1298

Abstract

Public programs and policies are designed and implemented with the aim of changing a certain behavior or aspect of a person's life, such as their education, income, diet, etc. To achieve its objectives, such implementation is based on theory of change. However, determining whether these public programs or policies achieve the expected impacts or results is a task that is often not addressed by the implementers of the programs or policies. Impact evaluation of public policies is a fundamental action within public management. It allows information to be obtained on the impacts that these policies or programs have on the beneficiary population and, thus, make it possible to study the option of expanding the scope, modifying, suggesting changes or stopping execution. In this systematic review, 22 impact evaluation studies were examined. The temporality of the studies ranges from 1983 to 2017. One of the main findings refers to the importance of estimating a credible counterfactual that allows an adequate identification of the impact of the application of a certain public policy. Another aspect to highlight is that the methodology to evaluate a certain public policy depends on the data available and the characteristics of the intervention. Finally, the econometric models used to estimate the impacts are by no means rigid models. Much depends on the ability and creativity of the researcher to formulate a methodological proposal that meets the planned objective.

Keywords: impact evaluation, selection bias, matching, difference-in-differences, discontinuity, treatment effect.

1. Introducción

Samuelson y Nordhaus (2006, p. 4) definen la economía como «el estudio de la manera en que las sociedades utilizan los recursos escasos para producir mercancías valiosas y distribuirlas entre los diferentes individuos». En esa línea, una parte fundamental del trabajo que realiza un gobierno es el diseño, gestión y evaluación de sus intervenciones, específicamente, de sus programas y políticas públicas.

La evaluación de las políticas públicas es una acción fundamental dentro de la gestión pública. Ella permite obtener información sobre los impactos que tienen dichas políticas o programas en la población beneficiaria y así poder estudiar la posibilidad de ampliar el alcance, modificar, sugerir cambios o, en su defecto, detener la ejecución. Holland y Rubin (1988) mencionan que, durante muchos años, la única contribución de los estadistas fue medir el «efecto causal», pero no explicarlo. Esto se acuñó en una conocida frase: «los estadistas pueden establecer correlación, pero no causalidad»; debido a que la determinación de la causalidad, sin una adecuada comprensión de los mecanismos causales, sería simplemente una correlación. Es aquí donde podemos situar las metodologías de evaluación de impacto. Estas permiten aprovechar la información (microdatos) para identificar relaciones de causalidad generadas por intervenciones públicas, llámense programas o políticas públicas.

Mejorar la efectividad de las políticas públicas requiere del conocimiento de los efectos y de las relaciones de causalidad atribuibles a estas. Conocer las relaciones de causalidad permite a los tomadores de decisiones identificar los programas o políticas públicas que sean más efectivas (menor inversión, mayores resultados).

Este artículo de revisión presenta una mirada a la literatura teórica de los principales métodos de evaluación de impacto que se utilizan en la actualidad para evaluar políticas públicas. Sin dejar completamente de lado algunas consideraciones formales, la redacción busca presentar un estilo amigable para quien pretenda introducirse en estos temas no se desanime por la dureza con la que se suele escribir la econometría.

2. Metodología

Este artículo presenta una revisión sistemática de la literatura. De acuerdo con Letelier, Manríquez y Rada (2005), «las revisiones sistemáticas son resúmenes claros y estructurados de la información disponible orientada a responder una pregunta específica». En este artículo, la pregunta a la que se pretende responder es: ¿cuáles son las principales bases teóricas de los métodos de evaluación de impacto? Para responder a esta pregunta, se realizó una amplia labor de recolección y análisis de los principales *papers* relacionados a la evaluación de impacto de programas y políticas públicas.

La primera tarea operativa que se realizó fue identificar las bases de datos, en las cuales se buscó la literatura teórica de los métodos de evaluación de impacto de políticas públicas correspondientes, para poder establecer el estado del arte de la materia en cuestión sobre la base de fuentes confiables. En la tabla 1 se muestran las bases consultadas.

Tabla 1
Bases de datos consultadas

Nombre	Dirección Url
Scientific Electronic Library Online	https://scielo.org
Red de Revistas Científicas de América Latina y el Caribe, España y Portugal	https://www.redalyc.org
Scopus	https://www.scopus.com
Researchgate	https://www.researchgate.net
Google Scholar	https://scholar.google.com

Luego de haber seleccionado las bases de datos, se elaboró una lista de palabras clave que guiaron la búsqueda de información para la elaboración del artículo de revisión. Estas palabras claves están relacionadas directamente con el objetivo y la pregunta de investigación. En algunos casos se utilizaron sinónimos de las palabras claves. La bibliografía sobre evaluación de impacto es amplia, por lo que se optó por seleccionar solo los documentos que contenían bases teóricas que permitan conceptualizar las técnicas de evaluación de impacto.

Otro aspecto que se consideró para la selección de los trabajos de investigación fue la temporalidad de estos, ya que estas técnicas se han ido perfeccionando con el tiempo, desde establecer sus bases a partir de 1980 hasta la actualidad. En un primer momento, se previó la revisión solo de trabajos en inglés, dado que los autores que establecieron las bases teóricas de la evaluación de impacto publicaron en ese idioma. Finalmente, se optó por revisar también trabajos publicados en español. En la tabla 2 se muestran los criterios utilizados para seleccionar el material a revisar.

Tabla 2
Criterios de selección del material

Selección	Rechazo
Es un trabajo de investigación teórico.	Es un trabajo de investigación empírico.
Fue elaborado después de 1980.	Fue elaborado antes de 1980.
Se encontró el documento original completo.	Documento al que no se puede acceder al texto completo.
Debe referirse específicamente a alguno de los métodos de evaluación de impacto.	-

Luego de haber discriminado los trabajos de investigación de acuerdo con los criterios de selección, se pasó a clasificar la bibliografía en cuatro categorías de análisis. La primera categoría está conformada por los trabajos de investigación relacionados con el método de «Diferencias en Diferencias». La segunda categoría está conformada por los trabajos de investigación relacionados con el método de «Propensity Score Matching». La tercera categoría está conformada por los trabajos de investigación relacionados con el método de «Randomized Control Trials». Finalmente, la cuarta categoría está conformada por los trabajos de investigación relacionados con el método de «Regresión Discontinua». Por último, se procedió a revisar las publicaciones e investigaciones teóricas sobre los principales métodos de evaluación de impacto.

3. Resultados

Para alcanzar el objetivo de esta revisión sistemática se revisaron 22 trabajos de investigación teóricos sobre evaluación de impacto de los 48 encontrados con las palabras clave. En la tabla 3 se presenta la relación de trabajos que fueron revisados y que sirvieron para desarrollar el presente artículo.

Tabla 3

Trabajos de investigación revisados

Nº	Título	Año	Autor	Objetivo	Resultados	Metodología
1	Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania	1994	David Card Alan B. Krueger	Comparar los cambios en el empleo, luego de un cambio en el salario mínimo.	Un aumento del salario mínimo, no disminuye el nivel de empleo.	Diferencias en Diferencias
2	Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*	2005	Alberto Abadie	Encontrar una estrategia de identificación para lidiar con la falta del cumplimiento del «supuesto de tendencias paralelas».	Utilizar estimadores de diferencias en diferencias semiparamétricos son útiles para afrontar la ausencia del «supuesto de tendencias paralelas».	Diferencias en Diferencias
3	Difference in Difference Meets Generalized Least Squares: Higher Order Properties of Hypotheses Tests	2007	Jerry Hausman Guido Kuersteiner	Encontrar una estrategia de identificación para lidiar con modelos que muestran alta correlación aplicando MCO.	Implementando MCGF más una expansión de Edgeworth se mejora la estimación realizada por MCO.	Diferencias en Diferencias
4	Matching on The Estimated Propensity Score	2009	Alberto Abadie Guido Imbens	Derivar la gran distribución muestral de los estimadores del PSM.	La estimación del puntaje de propensión afecta a la gran distribución muestral. Ajustan la varianza muestral para calcular un mejor puntaje de propensión.	Propensity Score Matching

5	Difference-in-Differences Estimation	2007	Guido Imbens Jeffrey Wooldridge	Revisar los métodos estándar de Diferencias en Diferencias.	Revisan tres métodos de Diferencias en Diferencias. Revisando los potenciales problemas de inferencia cuando se tiene más de dos grupos y periodos. Asimismo, revisan una aproximación no paramétrica y la construcción de grupo de control sintético.	Diferencias en Diferencias
6	The Estimation of Causal Effects by Difference-in-Difference Methods	2011	Michael Lechner	Revisar los métodos de estimación de Diferencias en Diferencias enfocándose en el efecto del tratamiento.	Es uno de los métodos más sencillos e intuitivos para evaluar políticas públicas. Una de las principales desventajas es que se debe cumplir el «supuesto de tendencias paralelas».	Diferencias en Diferencias
7	Inference in Difference-in-Differences with Few Treated Groups and Heteroskedasticity	2017	Bruno Ferman Cristine Pintoz	Encontrar una estrategia para lidiar con una evaluación de impacto con pocos grupos de tratados y presencia de heterocedasticidad.	Utilizando MCGF se pueden corregir ambos problemas y obtener un estimador consistente.	Diferencias en Diferencias
8	The Central Role of the Propensity Score in Observational Studies for Causal Effects	1983	Paul R. Rosenbaum	Mostrar cómo el PSM contribuye a eliminar el sesgo de selección de las variables observadas.	Propone un cuerpo teórico para la implementación del PSM.	Propensity Score Matching

9	Matching as an Econometric Evaluation Estimation	1998	James J. Heckman Hidehiko Ichimura Petra Todd	Desarrollar el método del PSM como un estimador de evaluación econométrica.	Si se utiliza la probabilidad o un puntaje conocido para emparejar, la varianza no mejora utilizando una u otra técnica.	Propensity Score Matching
10	A Primer for Applying Propensity-Score Matching	2010	Carolyn Heinrich Alessandro Maffioli Gonzalo Vázquez	Mostrar cómo se implementa un PSM para evaluar una política pública.	El PSM sirve mejor cuando se tiene sesgo con variables observables.	Propensity Score Matching
11	Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching	2010	Jyotsna Jalan Martin Ravallion	Evaluar el impacto de un programa de trabajo social en Argentina sobre el ingreso neto.	Aplicando PSM encuentran que los beneficiarios del programa, en promedio, tienen mayores beneficios.	Propensity Score Matching
12	Mostly Harmless Econometrics: An Empiricists Companion	2008	Joshua D. Angrist Jörn-Steffen Pischke	Desarrollan los principales conceptos relacionados a los RCT, DiD e IV.	Muestran la forma de implementar las distintas estrategias de evaluación de impacto.	Propensity Score Matching Diferencias en Diferencias
13	The Econometrics of Randomized Experiments	2016	Susan Athey Guido W. Imbens	Mostrar los principales métodos estadísticos para diseñar y analizar experimentos aleatorios.	Recomiendan utilizar la aleatorización en grupos pequeños y luego una segunda aleatorización.	Randomized Control Trials

14	The Case for Randomized Field Trials in Economic and Policy Research	1995	Gary Burtless	Examinar la justificación de la aplicación de los RCT en economía.	El experimento es la mejor combinación de fiabilidad, practicidad y rentabilidad en relación con otros tipos de evaluación.	Randomized Control Trials
15	Understanding and Misunderstanding Randomized Controlled Trials	2018	Angus Deaton Nancy Cartwrigth	Mostrar los pros y los contras de utilizar un RCT para evaluar políticas o programas.	El RCT arroja una estimación no sesgada pero siempre dentro de la muestra elegida. La validez externa es difícil de extrapolar.	Randomized Control Trials
16	The Regression Discontinuity Design: Theory and Applications	2008	Guido W. Imbens Thomas Lemieux	Mostrar cómo se implementan los métodos de RD.	Presentan conceptos básicos del RD, análisis gráfico, estimaciones y análisis de sensibilidad.	Regression Discontinuity
17	Incentive Effects of Social Assistance: A Regression Discontinuity Approach	2007	Thomas Lemieux Kevin Milligan	Evaluar el impacto de una política de asistencia social en el mercado de trabajo en Quebec.	Se encontró evidencia de que una mayor asistencia social reduce el empleo.	Regression Discontinuity
18	Regression Discontinuity Designs in Economics	2010	David S. Lee Thomas Lemieux	Mostrar la teoría básica para aplicar una evaluación de impacto utilizando una RD.	Muestran los principales conceptos del RD, identificación, interpretación y problemas de estimación.	Regression Discontinuity

19	Regression Discontinuity Design: Theory and Applications	2017	Matias D. Cattaneo Juan Carlos Escanciano	Presentar aspectos teóricos y prácticos para aplicar una RD.	Muestras metodológicas fundamentales como la identificación e interpretación, implementación, pruebas e inferencia.	Regression Discontinuity
20	Identification and Estimation of Treatment Effect with a Regression Discontinuity Design.	2021	Jinyong Hahn Petra Todd Wilber Van der Klaauw	Presentan una forma de evaluar los impactos utilizando una RD de forma no paramétrica.	Consideran el problema de la identificación con una RD Fuzzy y una RD Sharp.	Regression Discontinuity
21	A Practical Guide to Regression Discontinuity. University of Michigan	2012	Pei Zhu Marieé André Somers Howard Boom	Mostrar cómo aplicar en la práctica una RD.	Muestran cómo hacer el análisis gráfico, estimación paramétrica y no paramétrica, intervalo de validez e impacto y precisión.	Regression Discontinuity
22	La evaluación de impacto en la práctica, Segunda edición	2017	Paul J. Gertler Sebastián Martínez Patrick Premand Laura B. Rawlings Christel M. J. Vermeersch.	Mostrar los principales métodos de evaluación de impacto de políticas públicas.	Muestran la metodología del PSM, DiD, RCT y RD.	Propensity Score Matching Diferencias en Diferencias Randomized Control Trials Regression Discontinuity

4. Principales hallazgos

4.1. Diferencias en Diferencias

Este método de evaluación se utiliza cuando la «asignación del tratamiento», es decir, la asignación de los individuos al grupo que recibirá la intervención propuesta (tratamiento), no es muy clara para el investigador. Muchas veces las intervenciones o programas que se pretenden evaluar no tienen una regla clara de elección de beneficiarios; en estos casos, la evaluación de impacto se puede realizar a través del método de Diferencias en Diferencias. Este método consiste, básicamente, en medir el cambio en las variables de resultado entre un grupo que recibe la intervención (tratamiento) y un grupo que no recibe la intervención (control).

El hecho de no poder determinar claramente cuál fue la regla de decisión para elegir a los beneficiarios nos da indicios de que debemos utilizar un diseño cuasiexperimental, el cual, por definición, carece de distribución aleatoria (White y Sabarwal, 2014). Por lo tanto, se justificaría el uso de esta metodología. Este método asume que las variables de resultado, tanto en el grupo de tratamiento como en el grupo de control, tendrían la misma tendencia o seguirían el mismo rumbo antes del programa y después del programa, en ausencia de este. Esto se conoce como el supuesto de «tendencias paralelas». Para utilizar la metodología de Diferencias en Diferencias, es fundamental que este supuesto se cumpla. En otro caso, la diferencia entre los grupos de tratamiento y control no podría atribuirse al programa y los resultados serían sesgados.

Card y Kruger (1994) son considerados como autores del *paper* seminal sobre Diferencias en Diferencias. En este *paper*, los autores estudian el efecto de un cambio en la política de los salarios mínimos y su impacto en las horas trabajadas. Los autores identificaron que, en abril de 1992, el estado de Nueva Jersey dispuso la elevación del salario mínimo. Este pasó de $4.25 a $5.05 por hora. Para realizar esta evaluación, se tomó como unidad de análisis a los trabajadores de los restaurantes de comida rápida de Nueva Jersey (grupo de tratamiento) y a los trabajadores de restaurantes de comida rápida del estado vecino, Pensilvania (grupo de control), dado que la elevación del salario mínimo había sido establecido solo para Nueva Jersey.

Los autores utilizan la variación del salario entre el grupo de tratamiento y el grupo de control, antes de la entrada en vigencia de la política de salarios y después de la aplicación de esta, para medir su impacto en las horas trabajadas. Para alcanzar este objetivo, los autores utilizaron la siguiente especificación:

$$\Delta E_i = \alpha + bX_i + cNJ_i + \varepsilon_i$$

Donde:
ΔE_i = cambio en el empleo del tiempo 1 al tiempo 2 en el restaurante *i*
X_i = conjunto de características del restaurante *i*
NJ_i = variable *dummy* que toma el valor de 1 para los restaurantes en Nueva Jersey

Los autores utilizan algunos supuestos para realizar esta investigación. Uno de ellos es asumir que Pensilvania, por ser un estado vecino de Nueva Jersey, tiene condiciones económicas muy parecidas. A esto se suma que los restaurantes de comida rápida suelen tener el mismo formato en todos los estados, lo que facilita la comparación. Si bien no se mencionan específicamente estos hechos como el cumplimiento del supuesto principal de «tendencias paralelas», se asume que equivale a ello.

Uno de los principales hallazgos de este estudio fue que no se encontró evidencia de que el incremento del salario mínimo en el estado de Nueva Jersey redujera el empleo (horas trabajadas) en los restaurantes de comida rápida. Otro hallazgo fue que, independientemente de si se comparaba los restaurantes de Nueva Jersey que fueron afectados por el incremento del salario mínimo (tratamiento) con restaurantes de Pensilvania (control) o con otros restaurantes de Nueva Jersey que pagaban salarios altos y que no fueron afectados por la ley, se encontró que el salario mínimo incrementó el nivel de empleo.

Siguiendo a Fredriksson y Magalhães de Oliveira (2019), la formalización del método de Diferencias en Diferencias básico analiza los datos de dos grupos y dos periodos; estos datos son típicamente individuales. Los datos se pueden presentar como un panel o un corte transversal. Idealmente, la selección de los grupos de tratamiento y de control deben ser aleatorios. Con lo antes mencionado, el efecto del tratamiento quedaría de la siguiente manera:

$$DiD = (\bar{y}_{s = \text{treatment}, t = \text{after}} - \bar{y}_{s = \text{treatment}, t = \text{before}}) - (\bar{y}_{s = \text{control}, t = \text{after}} - \bar{y}_{s = \text{control}, t = \text{before}})$$

Donde y es la variable de resultado, el subíndice s representa al grupo y t representa el tiempo. Este es un ejemplo para un caso sencillo de dos grupos y dos periodos. Sin embargo, se puede generalizar para varios años y varios grupos. A continuación, se muestra el modelo 2x2 básico de Fredriksson y Magalhães de Oliveira (2019) y el modelo general.

Modelo 2x2: $$y_{ist} = A_s + B_t + BI_{st} + \varepsilon_{ist}$$

Modelo general: $$y_{ist} = A_s + B_t + cX_{ist} + dZ_{st} + BI_{st} + \varepsilon_{ist}$$

Lechner (2011) estudia los supuestos del método de Diferencias en Diferencias y menciona que el supuesto de tendencias comunes o tendencias paralelas establece que las diferencias en los resultados potenciales en el grupo de control no están relacionadas con la pertenencia al grupo de control o tratamiento, en el periodo previo al tratamiento. Ello implica que, en ausencia del tratamiento, ambos grupos tendrían la misma tendencia en las variables de resultado.

Fredriksson y Magalhães de Oliveira (2019) mencionan también la presencia de otro supuesto importante para este método, el supuesto de valor de tratamiento estable. Este supuesto implica que el tratamiento se brinda de manera que no se genere un efecto *spillover* (efecto que tiene el tratamiento en otros grupos que no fueron tratados) entre los grupos de tratamiento y de control.

Abadie (2005) menciona que la técnica de Diferencias en Diferencias, si bien es ampliamente utilizada para evaluar el efecto de las intervenciones públicas, presenta una dificultad que radica en que el supuesto de las tendencias paralelas es un supuesto de identificación muy fuerte. El autor propone un modelo que permite identificar el efecto del tratamiento para los tratados, aun cuando el supuesto de tendencias paralelas no se cumpla. Para esto propone una estrategia de dos pasos.

Uno de los principales problemas que enfrenta la evaluación de impacto es no poder estudiar a un mismo individuo en dos situaciones distintas (esto es, con tratamiento y sin tratamiento), por lo que se recurre a la figura del contrafactual (es lo que habría ocurrido, cuál habría sido el resultado de los participantes del programa, si no hubieran participado en el mismo), dado que es muy difícil encontrar dos individuos con características iguales. Sin embargo, es posible, a través de técnicas estadísticas, encontrar dos grupos de personas que en promedio sean muy parecidos (grupo de tratamiento y de control) de acuerdo con lo mencionado por White y Sabarwal (2014). La elección de los grupos de control es una parte muy importante en la evaluación de impacto, debido a que son estos los que hacen posible medir el efecto del tratamiento.

Abadie (2005) lleva la discusión sobre el uso del método de Diferencias en Diferencias un paso más adelante e introduce dos potenciales problemas al momento de estimar el efecto del tratamiento en los tratados (ATT). El primero de ellos es la heterogeneidad entre individuos de los grupos de tratamiento y control. Y el segundo es que no se cumpla el supuesto de tendencias paralelas.

Abadie muestra un modelo clásico de Diferencias en Diferencias y los problemas que enfrenta este modelo respecto a la heterogeneidad entre individuos y al incumplimiento del supuesto de identificación de tendencias paralelas:

$$Y(i,t) = \delta(t) + \alpha\, D(i,t) + \eta(i) + \nu(i,t)$$

Luego, para corregir el problema de heterogeneidad entre individuos, puede agregarse a la ecuación clásica un conjunto de covariables *X*:

$$Y(i,t) = \mu(t) + X(i)'\pi(t) + \tau\, D(i,1) + \delta t + \alpha\, D(i,t) + \varepsilon(i,t)$$

Donde «$X(i)'\pi(t)$» es un conjunto de covariables que representan la heterogeneidad en la dinámica de los resultados a través del tiempo. De acuerdo con Angrist y Pischke (2008), esto contribuiría a disminuir la varianza de las estimaciones calculadas del coeficiente de regresión.

Para corregir el problema de incumplimiento del supuesto de identificación de tendencias paralelas, el autor propone utilizar «dos simples pasos». El primer paso se basa en un esquema de ponderación a través de puntajes de propensión para identificar el grupo de control. Este esquema, como se verá más adelante con mayor detalle, plantea estimar la probabilidad de que cada individuo de la muestra sea elegido como parte del tratamiento a partir de sus características observables. Esta probabilidad (puntaje de propensión) permite emparejar a individuos muy parecidos estadísticamente. Siguiendo a Abadie (2005), esto se puede conseguir a través de un modelo logit o probit. El segundo paso consiste en aplicar el método de Diferencias en Diferencias para identificar el ATT.

Esta metodología puede ser útil para identificar el ATT cuando las características observables de los grupos tratados y no tratados difieren entre ambos.

Hausman y Kuersteiner (2007) analizan la estimación y la inferencia en modelos econométricos de Diferencias en Diferencias. Los autores identifican que, cuando los modelos basados en Mínimos Cuadrados Ordinarios (MCO) presentan autocorrelación (correlación serial), los estimadores son ineficientes y pueden llevar

a cometer errores de estimación. Para obtener estimadores eficientes, los autores proponen el uso de Mínimos Cuadrados Generalizados Factibles (MCGF), acompañado de una expansión de Edgeworth.

De acuerdo con los autores, la implementación de los Mínimos Cuadrados Generalizados Factibles (MCGF) muchas veces se ve obstaculizada por la poca cantidad de observaciones en el corte transversal (es decir, se tiene un N pequeño), lo que dificulta una adecuada estimación. Para corregir esto, analizan las propiedades de una muestra pequeña de las pruebas basadas en MCDF, aplicando una expansión de Edgeworth (Rothenberg, 1988, citado en Hausman y Kuersteiner, 2007) de orden superior que les permite construir una versión de la prueba con corrección de tamaño.

Hausman y Kuersteiner (2007) encuentran que, aplicando una expansión de Edgeworth, los estimadores calculados son significativamente más poderosos que el de Mínimos Cuadrados Ordinarios y Mínimos Cuadrados Ordinarios Robustos cuando los datos presentan alta autocorrelación (correlación serial).

Cuando la unidad de análisis es la persona, una empresa o alguna unidad económica individual en datos de corte transversal, la presencia de heterocedasticidad es un problema bastante común. Uno de los supuestos de MCO es la homocedasticidad o, dicho en otras palabras, la ausencia de heterocedasticidad. El incumplimiento de este supuesto llevaría a obtener estimadores ineficientes, es decir, ya no serían de varianza mínima. Ferman y Pinto (2015) muestran que los métodos de inferencia usados en Diferencias en Diferencias podrían no tener un buen desempeño frente a grupos tratados pequeños o frente a la presencia de heterocedasticidad y a la presencia de autocorrelación.

Para tratar el problema de la heterocedasticidad, Ferman y Pinto (2015) asumen que la heterocedasticidad está dada por el número de observaciones en cada grupo. Con este supuesto, reescalan los residuos del grupo de control usando la estructura de la heterocedasticidad, de forma que luego puedan usar esta información para estimar la distribución de los errores del grupo de control. Al igual que en Hausman y Kuersteiner (2007), estos autores proponen la implementación de los Mínimos Cuadrados Generalizados Factibles (MCGF) para manejar la presencia de autocorrelación.

4.2 Propensity Score Matching (PSM)

Al igual que el método de Diferencias en Diferencias, el Propensity Score Matching (PSM), o Emparejamiento por Puntajes de Propensión, es un método cuasiexperimental, que consiste en construir un grupo de control sintético a partir de observaciones que no hayan participado o recibido el tratamiento. En su concepción más simple, este método intenta encontrar, a partir de variables observables de los individuos, una unidad no tratada con características similares a una unidad tratada, con el objetivo de poder compararlas en sus variables de resultado. El *paper* seminal sobre PSM fue el de Rosembaum y Rubin (1983), en el cual definen el *matching* como «un método de muestreo, que toma valores de una reserva potencial de unidades no tratadas para construir un grupo de control, en el que la distribución de las covariables de este grupo es similar a las del grupo de tratamiento» (p. 48).

Rosembaum y Rubin (1983) definen el efecto causal como la diferencia entre el esperado de

la variable de resultado si recibió el tratamiento r_1, menos el esperado de la variable de resultado si no recibió el tratamiento r_0, como se expresa a continuación:

$$E(r_1) - E(r_0)$$

Rosembaum y Rubin (1983) también señalaron dos condiciones o supuestos que se deben cumplir. La primera condición, $(y_1,y_0) \perp Z|X$, exige que la asignación del tratamiento sea independiente de los resultados potenciales condicionados a las covariables de referencia observadas. La segunda condición, $0<P(X)<1$, exige que cada sujeto tenga una probabilidad distinta de cero de recibir cualquiera de los tratamientos, de esta última se desprende el denominado «soporte común» (la región en la que se intersecan las probabilidades de las unidades tratadas y no tratadas).

Para hallar el efecto causal, dado que no se puede observar a la misma unidad cuando recibe el tratamiento y cuando no lo recibe, Rosembaum y Rubin (1983) proponen construir un grupo de control «artificial», con la ayuda de un puntaje de propensión que resume todas las características observables y permite emparejarlos. Sin embargo, señalan la importancia de tener un adecuado balanceo que permita comparar los grupos de forma más significativa. Para calcular el puntaje de propensión (probabilidad de ser asignado al tratamiento), Rosembaum y Rubin proponen el uso de un modelo logit. Ellos demostraron que, si la asignación del tratamiento es *strongly ignorable* (*unconfounded* o no confundida por terceras variables), el condicionamiento en la puntuación de propensión permite obtener estimaciones no sesgadas de los efectos promedio del tratamiento.

El emparejamiento propuesto por Rosembaum y Rubin (1983) pretende encontrar un individuo y_0 que sea comparable a un individuo y_1. Este emparejamiento está basado en sus características observables. Heckman, Ichimura y Todd (1998) mejoran el emparejamiento presentando un modelo basado en una función de densidad tipo kernel, en el cual el emparejamiento no solo identifica al vecino más cercado (individuo y_0 más parecido al individuo y_1), sino que construyen emparejamientos utilizando a todos los individuos de la muestra y reduciendo la ponderación de aquellos individuos que se encuentren más distantes.

El sesgo de selección surge porque solo se puede observar a un individuo en alguno de los dos casos: si recibe el tratamiento o si no recibe el tratamiento, dadas algunas características *X* observables. Al igual que en Rosembaum y Rubin (1983), se mencionan los supuestos de independencia condicional y de soporte común. Adicionalmente, en el trabajo de Heckman et al. (1998) se incluye un tercer supuesto, $y_0 \perp D|X$, el cual exige que la probabilidad de no recibir el tratamiento sea independiente de los resultados potenciales condicionados a las covariables de referencia observadas. Así, los autores relajan el supuesto de independencia condicional con el objetivo de construir el contrafactual y_0, dadas algunas características *X* observables.

El primer paso para la estimación de los efectos de aplicar un determinado tratamiento, bajo el uso de la metodología del PSM, es el cálculo de los puntajes de propensión (*propensity score*). Como señalan Abadie e Imbens (2009), el cálculo de los puntajes de propensión afecta los estimadores del PSM. Por este motivo, Abadie e Imbens (2009) proponen un método para corregir la varianza asintótica de los estimadores del PSM. En el mismo sentido, Schneeweiss

et al. (2009), enfocados principalmente en obtener un adecuado puntaje de propensión que permita, a su vez, un adecuado emparejamiento y por lo tanto una buena estimación, señalan que, en la práctica, la identificación de las covariables que permitirán el cálculo del puntaje de propensión puede ser todo un reto para el investigador, debido a la gran cantidad de covariables que pueden existir. Es importante recordar que las covariables deben estar relacionadas al tratamiento y al resultado y deben ser adecuadamente identificadas. Schneeweiss et al. (2009) desarrollan un algoritmo de varios pasos que permite empíricamente identificar covariables candidatas, priorizar covariables e integrarlas en un modelo de ajuste de confusión basado en puntajes de propensión.

Como todo método, el PSM tiene ventajas y desventajas. Una de las ventajas es que se puede utilizar para medir el efecto de una política cuando esta ya se ejecutó. Otra de las ventajas es que no requiere de aleatorización para medir el efecto de la política o programa. En cuanto a las desventajas, una de ellas es que el cálculo de los puntajes de propensión se realiza solo sobre variables observables, dejando de lado las variables no observables que podrían influir en los resultados. Otras de las desventajas es que, para lograr un adecuado cálculo de los puntajes de propensión, se requieren los datos de todas las variables observables que puedan influir en la elección del tratamiento o en el resultado, es decir, se necesita una gran cantidad de información.

4.3 Randomized Control Trial (RCT)

El método de asignación aleatoria (RCT, por sus siglas en inglés, Randomized Control Trial), es un método de evaluación de impacto que se asemeja a la realización de un sorteo, en el que la asignación al tratamiento es completamente aleatoria, asegurando de esta manera que cada individuo tenga la misma probabilidad de ser elegido. Esta asignación resulta ser la forma más justa e imparcial para destinar recursos escasos, así como para evaluar el impacto del programa (Gertler, Martínez, Premand, Rawlings y Vermeersch, 2017). Una evaluación bajo esta metodología es aplicable cuando la demanda (grupo de potenciales beneficiarios) supera ampliamente a la oferta del programa, o cuando los recursos del programa son limitados o cuando existen factores éticos que se contraponen al tipo de asignación elegida por un determinado programa.

De acuerdo con Angrist y Pischke (2008) la asignación aleatoria al tratamiento resuelve el problema de selección, ya que esta asignación aleatoria es independiente de los resultados potenciales. A esto se le conoce como el supuesto principal de un RCT, la independencia condicional:

$$(Y_1, Y_0) \perp\!\!\!\perp T$$

Bajo este supuesto, y siguiendo la notación que utilizan estos autores, se tiene que:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$$
$$= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]$$

Donde la independencia de Y_{0i} y D_i permite reemplazar $E[Y_{0i} \mid D_i=1]$ por $E[Y_{0i} \mid D_i=0]$, y, dada la aleatorización, la ecuación se simplifica de la siguiente manera:

$$E[Y_i \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] = E[Y_{1i} - Y_{0i} \mid D_i=1]$$
$$\text{ATE} = \text{ATT} = \text{ATU} = E[Y_{1i} - Y_{0i}]$$

Esto permite eliminar el sesgo de selección. Asimismo, dada la aleatorización, la diferencia en las variables de resultados entre los grupos de tratamiento y control captura el promedio del efecto causal. Como mencionan los autores, esto no deja libre de problemas a la evaluación, pero sí resuelve el problema más importante en las evaluaciones empíricas. Cuando se utiliza un RCT, en teoría, se deberían construir dos grupos (tratamiento y control) altamente similares, siempre y cuando se tenga un número grande de unidades.

Gertler et al. (2017) muestran que se deben realizar dos pasos para una adecuada asignación. El primer paso consiste en elegir una muestra de una población de unidades elegibles, a través de un sorteo (aleatorización). Esto permitirá lograr la validez externa de la evaluación. El segundo paso consiste en construir los grupos de tratamiento y control de la muestra previamente elegida. Esto también se debe realizar a través de un sorteo (aleatorización). Esto permitirá lograr la validez interna de la evaluación.

Shadish, Cook y Campbell (2002) señalan que un estudio presenta validez interna si las covarianzas observadas entre un tratamiento y su resultado reflejan una relación causal considerando todas las variables que fueron utilizadas. La validez interna se refiere a la capacidad de un estudio de estimar efectos causales dentro de la población de estudio. Un RCT bien ejecutado resuelve el problema de la validez interna. Asimismo, Shadish et al. (2002) mencionan que un estudio presenta validez externa cuando los resultados se pueden generalizar al conjunto de la población de unidades elegibles.

Para hallar el ATE (efecto promedio del tratamiento) usando un RCT, Athey e Inbems (2016)

señalan que en mínimos cuadrados ordinarios (MCO) se realiza una regresión como se muestra en la siguiente ecuación:

$$Y_i^{obs} = \alpha + \tau W_i + \varepsilon_i$$

Donde ε_i es el término de error, W_i es una variable dicotómica que asigna el tratamiento y τ es un estimador insesgado para el efecto causal promedio. Sin embargo, mencionan que uno de los supuestos fundamentales de un MCO es que el término de error ε_i es independiente o no está correlacionado con W_i y que este es difícil de evaluar. Además, mencionan que rara vez se interpreta más allá de solo una noción de que captura factores no observados que afectan al tratamiento. A este hecho debe sumarse que la aleatorización no implica la no correlación de ε_i con W_i. En general, existirá heterocedasticidad. Siguiendo con los autores, es necesario utilizar los errores estándar robustos de Eicker-Huber-White para obtener intervalos de confianza válidos.

Athey e Inbems (2016) señalan que se puede agregar covariables a la regresión clásica por MCO. Se asume que estas covariables no son afectadas por el tratamiento (covariables pretratamiento) y, además, no afectan a la asignación. Athey e Inbems (2016) mencionan que «la presencia de estas covariables deviene de tres motivaciones: i) mejorar la precisión de la estimación, ii) permiten incluir la heterogeneidad de los individuos, iii) permiten remover sesgo en comparaciones simples de media si la aleatorización no fue adecuada» (p. 27).

Si bien un RCT permite una evaluación muy creíble cuando esta se implementa adecuadamente, tiene varias desventajas. Burtless (1995) plantea que unos de los problemas que presenta un RCT son los elevados costos de implementación en comparación con otro tipo de evaluaciones, así como el tiempo que tiene que transcurrir entre la aplicación del tratamiento y la evaluación del impacto en las variables de resultado. Burtless (1995) también enfatiza sobre los problemas éticos que presenta este tipo de evaluaciones, sobre todo por la dificultad de elegir adecuadamente a los grupos de tratamiento y de control cuando la asignación al tratamiento puede perjudicar al tratado o al control. Burtless (1995) también menciona el problema de los individuos tratados que en el transcurso del tratamiento se retiran del programa, conocidos en inglés como los «attrition». Finalmente, un RCT debe implementarse antes del inicio del programa; esto quiere decir que con un RCT no se puede hacer evaluaciones a programas pasados. Estas son algunas de las principales desventajas que presenta la implementación de un RCT.

4.4 Regresión Discontinua (RD)

Los programas o políticas que se ejecutan, en ocasiones, utilizan un punto de quiebre o punto de inflexión para diferenciar entre los individuos que pueden ser parte de un tratamiento y los que no. Este punto de quiebre, normalmente, es un índice continuo, como la línea de la pobreza, la edad del individuo, el nivel de ingresos, etc., normalmente llamados «índices de elegibilidad». Un clásico ejemplo son los programas de apoyo a la reducción de la pobreza, los cuales suelen definir un umbral de pobreza a partir del cual se establece quiénes pueden ser beneficiarios y quiénes no, de un determinado tratamiento o programa. Un ejemplo específico de este tipo de programas es el Programa del Estado Peruano Pensión 65. Este programa entrega una subvención económica a adultos mayores, precisamente a los mayores de 65 años.

De acuerdo con Hahn, Todd y Klaauw (2001), «la regresión discontinua es un método de

evaluación cuasi-experimental con la característica de que la probabilidad de ser elegido o recibir el tratamiento cambia discontinuamente como función de uno o más variables subyacentes» (p. 201).

Siguiendo a Hahn et al. (2001), el modelo para hallar el efecto causal puede escribirse de la siguiente manera:

$$y_i = \alpha_i + x_i \beta_i$$

Donde x_i representa la asignación al tratamiento y β_i representa el efecto causal del tratamiento. Asimismo, menciona que existen dos tipos de diseños de regresiones discontinuas en la literatura: el diseño nítido (*sharp*) y el diseño difuso (*fuzzy*). En el diseño nítido, x_i es conocido y depende de una manera determinística de algunas variables observadas. En cambio, en el diseño difuso, x_i es una variable aleatoria que depende de algunas variables observadas. En otras palabras, en el modelo difuso, la asignación al tratamiento no es una función determinística de las variables observadas, dado que existen otras variables no observadas que determinan la asignación al tratamiento.

Para implementar un RD de forma adecuada, deben cumplirse algunas condiciones. La primera, de acuerdo con Hahn et al. (2001) es que el «índice de elegibilidad» debe ser continuo. Bajo este supuesto, el efecto del tratamiento se obtiene estimando la discontinuidad en el punto de quiebre, entre el promedio de las observaciones que están por encima del umbral y el promedio de las que están por debajo. La segunda condición refiere que la variable identificada como la variable que determina la participación no debe estar influenciada por el tratamiento. Esta se identifica al inicio del tratamiento y no puede ser modificada. La tercera condición señala que el punto de quiebre o umbral es independiente de la variable de clasificación y la asignación al tratamiento, y que está basada exclusivamente en la clasificación de los individuos y en el punto de corte o umbral. La cuarta condición es que la discontinuidad se debe dar en un solo punto del intervalo, lo que implica que las observaciones de un lado del umbral y del otro se tratan de manera similar, a excepción del tratamiento.

Según Cattaneo, Idrobo y Titiunik (2019), en un RD *sharp* todas las observaciones reciben un puntaje, la variable de clasificación es conocida y el tratamiento se asigna a las observaciones que están por encima del umbral (grupo de tratamiento) y no se asigna tratamiento a las que están por debajo del umbral (grupo de control). En el caso de un RD *fuzzy*, Zhu, Somers y Bloom (2012) señalan que este se implementa cuando el tratamiento es parcialmente determinado por la variable de clasificación y el umbral. Por ejemplo, cuando algunas observaciones del grupo de tratados no reciben el tratamiento o cuando algunas observaciones del grupo de control reciben el tratamiento.

Imbens y Lemiux (2008) sugieren que el análisis gráfico puede ser una parte fundamental para el análisis antes de la implementación de un RD. La naturaleza de un RD sugiere que el efecto del tratamiento se puede medir por el valor de la discontinuidad de la variable de resultado en un determinado punto de corte o umbral. Por lo tanto, si se elabora un gráfico de dispersión utilizando la variable de resultado y la variable explicativa, se podrá notar si existe una discontinuidad en la variable de resultado en el punto de corte.

Un RD bien implementado produce estimaciones muy fiables para programas que desean deter-

minar si amplían sus intervenciones o no. Una de las desventajas que presenta un RD, ya sea *sharp* o *fuzzy*, es que el efecto del tratamiento calculado es un efecto local, es decir, alrededor del umbral. No se podrán identificar los impactos para las observaciones que se encuentran más alejadas del umbral. Para mejorar la potencia de un RD, siguiendo a Gertler et al. (2017), el ancho de banda a utilizar debería ser lo suficientemente amplio para alcanzar el mayor número de observaciones, pero sin perder el nivel de discontinuidad que se genera en la variable de resultado.

5. Conclusiones

En esta revisión de literatura se evaluaron veintidós estudios sobre evaluación de impacto. La temporalidad de los estudios evaluados va desde el año 1983 hasta el año 2019. A partir de estos estudios, se ha realizado una revisión de la literatura teórica de los cuatro principales métodos de evaluación de impacto, tratando de mantener un *trade-off* entre la intuición de los métodos y la formalidad econométrica. De los cuatro métodos revisados, tres son métodos cuasiexperimentales (PSM, Diff in Diff y RD) y uno es un método experimental (RCT). La ventaja de los métodos cuasiexperimentales es que se pueden diseñar y aplicar antes, durante y después de la implementación de los programas o políticas públicas, mientras que el método experimental solo se puede aplicar desde el inicio del tratamiento.

Cada una de estas metodologías presenta ventajas y desventajas. Cada una de ellas está diseñada para evaluar un tipo de intervención, considerando sus características particulares. La idea de la evaluación de impacto es que estas metodologías se acomoden al tipo de programa o política que se pretende evaluar y no al revés. En otras palabras, el evaluador debe tener la capacidad de acomodar la metodología a la realidad. Asimismo, cada metodología presenta sus propios retos al momento de llevarlas a la práctica; es ahí donde entra la creatividad y la experiencia del investigador.

La importancia de la evaluación de impacto radica en la posibilidad de identificar las relaciones causales de los programas y políticas públicas. Esto permite generar evidencia robusta sobre qué programas o políticas públicas funcionan y cuáles no. Asimismo, permite identificar y sugerir mejoras a las intervenciones existentes, ampliar el alcance, o, en su defecto, cerrar determinas intervenciones. Esto contribuye a mejorar la efectividad del gasto público.

Cada vez se viene generando un mayor número de publicaciones (investigaciones, documentos de trabajo, etc.) sobre evaluación de impacto de políticas públicas, que en su mayoría son trabajos empíricos que se amoldan al tipo de intervención que se pretende evaluar. El hecho de tener un sinfín de intervenciones, cada una con sus características particulares, genera retos que deben ser resueltos por los evaluadores. Es aquí donde estas evaluaciones empíricas, sin tener como objetivo principal el generar conocimiento teórico, pueden generar nuevo conocimiento que permita a los nuevos evaluadores enfrentar los futuros retos con mejores herramientas. Por ejemplo, la pandemia generada por la COVID-19 supone un reto adicional con relación al recojo de información cuantitativa y cualitativa para realizar evaluaciones de impacto. De la misma manera, las variables han tenido comportamientos atípicos que deberán ser manejados adecuadamente, adaptando las metodologías de manera que se pueda aislar los resultados de las políticas, programas o proyectos evaluados durante esta pandemia.

6. Limitaciones

La principal limitación del presente artículo de revisión es que no se ha considerado la literatura empírica de los distintos métodos de evaluación de impacto. La literatura empírica incluye una gran variedad de documentos y cada uno de ellos presenta alguna particularidad relacionada con el tipo de intervención que se pretende evaluar; en un futuro artículo de revisión se podría organizar literatura empírica por cada método de evaluación.

Una segunda limitación del presente artículo es la ausencia de mayor detalle en cuanto a la formalidad econométrica. Ello se debe a que el objetivo de este artículo es servir de introducción a las metodologías de evaluación de impacto de manera general, lo que no impide que el lector pueda comprender claramente las formalidades básicas relacionadas con cada método.

Referencias bibliográficas

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72, 1-19.

Abadie, A. e Imbens, G. W. (2009). Matching on the estimated propensity score. Working Paper 15301. National Bureau of Economic Research. http://www.nber.org/papers/w15301

Angrist, J. D. y Pischke, J-S. (2008). *Mostly harmless econometrics: An empiricist's companion*.

Athey, S. e Imbens, G. W. (2016). The econometrics of randomized experiments. Papers 1607.00698, arXiv.org.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives*, *9*(2), 63-84.

Card, D. y Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772-793.

Cattaneo, M., Idroboy, N. y Titiunik, R. (2019). *Practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. y Escanciano, J. C. (2017). Regression discontinuity designs: Theory and applications. *Advances in Econometrics*, *38*.

Ferman, B. y Pinto, C. (2015). *Inference in differences-in-differences with few treated groups and heteroskedasticity*. Sao Paulo School of Economics – FGV.

Fredriksson, A. y Magalhães de Oliveira, G. (2019). *Impact evaluation using difference-in-differences*. Center for Organization Studies (CORS), School of Economics, Business and Accounting (FEA), University of São Paulo (USP), São Paulo, Brazil.

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B. y Vermeersch, C. M. J. (2017). *La evaluación de impacto en la práctica*, Segunda edición. Washington, DC: Banco Interamericano de Desarrollo y Banco Mundial. doi:10.1596/978-1-4648-0888-3.

Hahn, J., Todd, P. y Klaauw, W. V. (2001). Identification and estimation of treatment effect with a regression discontinuity design. *Econometrica*, *69*(1), 201-209.

Hausman, J. y Kuerteiner, G. (2007). *Difference in difference meets generalized least squares: Higher order properties of hypotheses tests*.

Heckman, J. J., Ichimura, H. y Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies,* 65, 261-294.

Holland, P. W. y Rubin, D. (1988). *Causal inference in retrospective studies.*

Imbens, G. W. y Lemieux, T. (2008). The regression discontinuity design —theory and applications. Special Issue, *Journal of Econometrics*, *142*(2), 611-614.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165-224.

Rosembaum, P. R. y Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Shadish, W. R., Cook, T. D. y Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Schneeweiss, S., Rassen, J., Glynn, R., Avorn, J., Mogun, H. y Brookhart, A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 2009; 20:512-522.

White, H. y Sabarwal, S. (2014). Diseño y métodos cuasiexperimentales. *Síntesis metodológicas: evaluación de impacto,* 8, Centro de Investigaciones de UNICEF, Florencia.

Zhu, P., Somers, M. A. y Boom, H. (2012). *A practical guide to regression discontinuity*. University of Michigan.

Fecha de recepción: 1 de diciembre de 2020
Fecha de aceptación: 14 de diciembre de 2020