MÉTODOS DE EVALUACIÓN DE POLÍTICAS PÚBLICAS

Joaquín ARTÉS (*)
Beatriz RODRÍGUEZ-SÁNCHEZ
Universidad Complutense de Madrid

Resumen

Las políticas públicas se diseñan para cumplir determinados objetivos. Las técnicas de evaluación de impacto causal permiten medir los efectos de las políticas públicas en las variables de interés de forma que pueda evaluarse hasta qué punto la política ha logrado sus objetivos. En este artículo se hace una revisión de los principales métodos de evaluación de impacto causal utilizados habitualmente en la investigación académica.

Palabras clave: evaluación, políticas públicas, inferencia causal.

Abstract

Public policies are implemented with the intention of achieving certain goals. The effects of such policies can be measured using various impact evaluation techniques, providing crucial information regarding the extent to which a policy has achieved such goals. In this paper, we review the main methods that are most frequently applied in academic research to evaluate the impact of public policies.

Keywords: evaluation, public policy, causal inference.

JEL classification: C50, H00.

I. INTRODUCCIÓN

LAS políticas públicas se diseñan para cumplir una serie de objetivos como pueden ser la mejora del rendimiento académico o la mejora de la salud, por citar algunos ejemplos. La evaluación de los efectos de las políticas públicas es fundamental para asegurar que la política cumple los objetivos para los que fue diseñada, para cambiar su diseño en caso de que no los cumpla y para asegurar que el dinero público se usa eficientemente.

El problema que surge a la hora de evaluar los efectos de una política pública es cómo diferenciar los cambios en la variable objetivo (*e. g.,* el rendimiento académico) que se deben a la política implementada de aquellos cambios que se deben a otros aspectos que también pueden influir en esa variable. Un análisis que no permita analizar la relación causa-efecto entre la política pública y la variable objetivo, o lo haga erróneamente, conducirá a conclusiones incorrectas y, en consecuencia, a decisiones erróneas.

En los últimos años, la literatura académica ha desarrollado numerosos métodos de evaluación de impacto de las políticas públicas que permiten determinar con gran precisión si un cambio en una variable objetivo es causado por una determinada política pública. Pese a su importancia, el análisis riguroso de los efectos que cabe atribuir a una política pública no es habitual en el ámbito de la gestión pública en España. Con objeto de contribuir a superar esta laguna, en este artículo describimos los principales métodos de evaluación de impacto causal que se utilizan habitualmente tanto en la literatura académica como en la práctica profesional. Comenzamos el trabajo con una descripción del problema analítico que representa la identificación de relaciones causales en el ámbito de las políticas públicas para, a continuación, hacer un repaso de los principales métodos estadísticos que permiten identificar relaciones causa-efecto.

II. EL PROBLEMA FUNDAMENTAL DE LA INFERENCIA CAUSAL EN LA EVALUACIÓN DE POLÍTICAS PÚBLICAS

Supongamos, a efectos ilustrativos, que queremos evaluar el efecto que supone recibir clases de refuerzo durante el último año de bachillerato en el rendimiento académico obtenido en el examen de acceso a la universidad. Una simple correlación entre acudir a clases de refuerzo y el rendimiento académico obtenido en el examen podría mostrar una relación negativa entre ambas variables: a más clases de refuerzo, peor rendimiento académico en el examen de acceso. Esto podría ser debido a que los alumnos que acuden a clases de refuerzo son sustancialmente distintos a aquellos que han optado por no acudir a clases de refuerzo, dado que, previsiblemente, estos últimos son mejores estudiantes. Obviamente, sería un error interpretar esta correlación de forma causal y concluir que las clases de refuerzo empeoran el rendimiento académico.

El uso de una metodología adecuada resulta crucial en casos como el mencionado y, en general, en cualquier análisis de políticas públicas, dado que permite identificar el efecto real de la política pública (*e. g.,* las clases de refuerzo), separando ese efecto del que pudieran tener otras variables relacionadas (*e. g.,* la capacidad previa de los estudiantes incluidos en la comparación) en la variable objetivo.

El problema que se plantea desde el punto de vista del análisis causal es que para cada individuo que incluyamos en nuestro análisis únicamente observaremos un estado de la naturaleza: o bien el individuo ha recibido clases de refuerzo en el momento de hacer el examen de acceso, o no las ha recibido. Este problema hace que no se pueda calcular directamente el efecto causal de esa variable para cada individuo concreto. Eso no impide, sin embargo, siempre que dispongamos de una muestra suficientemente grande de individuos, que podamos calcular el efecto causal que ha tenido el programa, en media, comparando el rendimiento de un grupo de individuos beneficiario del programa (*grupo tratado)* con el de un grupo de individuos no beneficiarios del mismo que sean lo suficientemente parecidos (*grupo de control*). La razón es que al comparar dos grupos de individuos que, en media, tienen características similares en todas las variables que podrían influir en la variable objetivo, excepto el tratamiento recibido, las diferencias en la variable objetivo que se observen solo podrían ser atribuidas al tratamiento, precisamente porque, en el resto de las características, ambos grupos serían iguales.

Si ambos grupos fuesen diferentes, al comparar la nota media obtenida en el examen de acceso a la universidad de cada grupo, esta diferencia vendría explicada, en parte, por el efecto que estas otras características tienen en el rendimiento académico. En este caso, la diferencia de notas medias entre los dos grupos podría descomponerse en dos efectos: el efecto del tratamiento y el efecto de estas otras características que diferencian a ambos grupos que, siguiendo la terminología habitual, podemos denominar sesgo de selección. Es decir, la existencia de dicho sesgo hace que la diferencia de medias no coincida con el efecto causal del tratamiento que queremos analizar.

El problema fundamental de la evaluación de políticas públicas consiste, en consecuencia, en encontrar un grupo contrafactual adecuado que permita estimar de forma precisa lo que habría ocurrido con la variable objetivo en ausencia de la política pública que ha afectado al grupo de tratamiento, de forma que se elimine el sesgo de selección.

III. EL ESTÁNDAR DE ORO: LA ALEATORIZACIÓN

Una forma ideal de asegurar que el grupo de control es estadísticamente idéntico al grupo tratado, excepto en el tratamiento recibido, consiste en aleatorizar *ex ante* qué individuos reciben el tratamiento y cuáles no. Este es el procedimiento utilizado habitualmente en el campo de las ciencias de la salud al evaluar la eficacia de un medicamento (*e. g.,* una vacuna). La forma habitual de medir el impacto causal de una vacuna consiste, en primer lugar, en seleccionar un grupo de población potencialmente susceptible de contraer una enfermedad. A continuación, se sortea (se aleatoriza) a qué individuos se les administra el medicamento y a cuáles no (o a cuáles se les administra un placebo). Pasado un tiempo, se comparan los porcentajes correspondientes a los individuos del grupo tratado y del grupo de control que han contraído la enfermedad contra la que la vacuna pretende ser efectiva. Una diferencia estadísticamente significativa en el porcentaje de contagios en ambos grupos puede ser interpretada de forma causal porque la aleatorización, junto con la inclusión de un número suficientemente grande de individuos en el análisis, asegura que el grupo tratado y el grupo de control son estadísticamente idénticos.

El uso de experimentos aleatorios (*RCT* o *randomized controlled trials*) en el ámbito de las ciencias sociales presenta dificultades añadidas dadas las características de los tratamientos a analizar. Sin embargo, las dificultades adicionales no han impedido que desde finales de los años noventa se haya generalizado su uso para investigar una amplia variedad de cuestiones (1). En este sentido, y a modo de ilustración, mencionaremos dos tipos de estudios representativos.

En primer lugar, la evaluación de programas piloto en el ámbito educativo o sanitario. Un programa piloto consiste en implementar una política pública a pequeña escala para poder evaluar sus efectos antes de aplicar el programa a gran escala. El programa STAR, en el estado de Tennessee, en Estados Unidos, es la referencia clásica en este tipo de estudios (véase Krueger,

1999). El programa STAR pretendía evaluar el efecto de la reducción del número de alumnos por aula en el rendimiento académico aleatorizando el número de alumnos por aula en los colegios que voluntariamente eligieron participar en el piloto. La aleatorización permitió demostrar de forma causal que la reducción del número de alumnos por aula supuso una mejora de 5 puntos porcentuales en el rendimiento académico de los alumnos que fueron asignados aleatoriamente a clases pequeñas en comparación con el grupo contrafactual. En España, sin embargo, apenas existen artículos académicos que muestren los resultados de una evaluación de políticas educativas utilizando ensayos aleatorios, aunque sí existe algún ejemplo en el ámbito de la economía de la salud. Un estudio reciente realizado en siete países europeos, entre los que se encuentra España, mostró el coste-efectividad (potenciales ahorros y mejores resultados en salud) de una intervención multimodal consistente en un programa de actividad física y consejos nutricionales en personas mayores de 70 años con limitaciones funcionales (Peña-Longobardo *et al*., 2021).

Un segundo tipo de estudios que resulta útil mencionar, por su relevancia, son los trabajos que utilizan la aleatorización como técnica para medir la existencia de discriminación en el ámbito laboral o educativo. Una de las referencias clásicas en este contexto es el estudio de Bertrand y Mullainathan (2004). En este trabajo se pretendía medir la existencia de discriminación racial en el mercado de trabajo de Estados Unidos. El tratamiento consistió en manipular el nombre que aparecía en el *curriculum vitae* (CV) enviado a distintas ofertas de trabajo publicadas en periódicos de Boston y Chicago, siendo el resto de las características del CV idénticas entre el grupo tratado y el grupo de control. En el grupo tratado el nombre que aparecía en el CV era típicamente afroamericano, mientras que el grupo de control incluía nombres comúnmente asociados con individuos de raza blanca. El estudio muestra que las solicitudes asociadas con nombres afroamericanos recibían un 50 por 100 menos de llamadas para realizar una entrevista en comparación con solicitudes idénticas asociadas a nombres de raza blanca. En España, este tipo de aleatorización se ha utilizado, por ejemplo, para estudiar la discriminación frente al colectivo homosexual (Díaz-Serrano y Meix-Llop, 2016), o para estudiar la discriminación por razón de género (Fernández-Cornejo, 2011).

IV. EXPERIMENTOS NATURALES

En los dos tipos de análisis mencionados anteriormente se elimina el sesgo de selección mediante la aleatorización *ex ante* de tratamiento. La aleatorización *ex ante*, sin embargo, no puede llevarse a cabo para evaluar la gran mayoría de políticas públicas. Dos son las razones: o bien porque no se previó la realización de un piloto antes de la implementación de la política pública por su elevado coste o porque la naturaleza del tratamiento impide su aleatorización. En algunos casos, sin embargo, también puede llevarse a cabo una evaluación causal apoyada en la aleatorización que se produce involuntariamente, de forma «natural», por la forma en la que se llevó a cabo la política pública o por un fenómeno azaroso que permite identificar individuos que aleatoriamente han recibido o no un determinado tratamiento de interés. A este tipo de estudios se les denomina experimentos naturales.

En el ámbito de las ciencias sociales y del análisis de políticas públicas ha habido muchos ejemplos de experimentos naturales a lo largo de las últimas décadas. Uno de los ejemplos recientes más conocido es el análisis de los efectos de Medicaid, seguro público de salud estadounidense para individuos de rentas bajas (véase Finkelstein *et al.,* 2012). En el año 2008, en Oregón, se abrió de forma transitoria la posibilidad de que nuevos individuos que no tenían suscrito un seguro privado de salud pudieran beneficiarse del programa Medicaid. La respuesta fue una avalancha de nuevas solicitudes de forma que el estado de Oregón, al no poder cubrir todas las solicitudes, decidió sortear entre ellas a quienes tendrían finalmente acceso al programa. La decisión política de sortear qué individuos serían finalmente beneficiarios es equivalente a un ensayo aleatorio, aunque no se hubiese planificado previamente. Esa decisión permitió posteriormente estudiar los efectos causales de tener seguro de salud en numerosas cuestiones de interés como la frecuencia de utilización de servicios sanitarios, la salud física o mental, el empleo, el salario recibido o incluso la participación política (2).

En España, también hay circunstancias institucionales que dan lugar a aleatorizaciones naturales de interés para las ciencias sociales y las políticas públicas. Mencionaremos, a efectos ilustrativos, el análisis de los sesgos de género en la selección de los funcionarios públicos realizado por Bagues y Esteve-Volart (2010). Los distintos tribunales de evaluación de los candidatos a distintos cuerpos

funcionariales de la Administración española, desde el año 1987 al año 2007, difieren en la composición de género y, en particular, en el porcentaje de mujeres que lo forman. Los candidatos son asignados a los distintos tribunales a través de un sorteo que determina qué candidatos actuarán en primer lugar en las pruebas orales. Este sistema produce una aleatorización natural ya que los candidatos no son asignados a un determinado tribunal por su género, sino por la letra de su apellido. Este hecho hace que, *de facto*, tenga lugar una aleatorización de los candidatos a tribunales compuestos por más o menos mujeres. Bagues muestra un efecto causal de la composición de género del tribunal en la probabilidad de que candidatos de un género u otro aprueben la oposición, teniendo las mujeres menos probabilidades de aprobar, *ceteris paribus*, cuanto mayor es el porcentaje de mujeres en el tribunal.

En definitiva, el análisis causal de los efectos de determinadas políticas públicas puede analizarse de forma muy precisa y sencilla desde el punto de vista estadístico (una diferencia de medias es suficiente en la mayoría de los casos) cuando se realiza una aleatorización del tratamiento de interés entre individuos similares. Obviamente, el problema al que se enfrentan los investigadores en la mayor parte de situaciones consiste en que el grupo que recibe el tratamiento que se quiere analizar y los potenciales grupos de control no están aleatorizados, por lo que suele existir un sesgo de selección potencialmente muy severo. A continuación, analizamos una serie de técnicas estadísticas que permiten afrontar esos problemas.

V. MÉTODOS PARA CONTROLAR VARIABLES OBSERVABLES

La forma más directa de eliminar los sesgos de selección producidos por las diferencias que existen entre el grupo que recibe el tratamiento que queremos analizar y los potenciales grupos de control es utilizar un modelo estadístico que tenga en cuenta esas diferencias. Las dos técnicas que se utilizan más habitualmente para controlar por las diferencias no debidas al tratamiento entre el grupo tratado y el grupo de control son los modelos de regresión y los modelos de emparejamiento (o modelos *matching*).

1. Modelos de regresión

Los modelos de regresión siguen siendo el principal instrumento empírico con el que cuentan los investigadores para analizar los efectos de las políticas públicas en ausencia de ensayos aleatorios o experimentos naturales. Lo usual es disponer de datos observacionales, recogidos una vez aplicada una determinada política pública en la que no se realizó ninguna aleatorización previa del tratamiento, por lo que los individuos que participan en el programa que se quiere analizar suelen ser distintos de aquellos que no participan en el programa. En estos casos, una regresión multivariante puede solventar el problema fundamental de la inferencia causal y dar lugar a estimaciones causa-efecto de las políticas públicas. Sin embargo, la inferencia a partir de una regresión únicamente será causal siempre y cuando se incluyan como controles en la regresión todas aquellas variables de control que son susceptibles de afectar a la variable de interés y que diferencian a los individuos del grupo tratado de los individuos del grupo de control. En esos casos decimos que se cumple la hipótesis de la independencia condicional.

La independencia condicional implica que cuando interpretamos una regresión mínimo-cuadrática (MCO) estamos asumiendo que, una vez que se controlan los efectos de las variables de control en la variable de interés, las diferencias en la variable objetivo entre el grupo tratado y el grupo de control se deben únicamente al tratamiento, como en un experimento aleatorizado. Podemos recurrir de nuevo al ejemplo de las clases de refuerzo para aclarar cómo podría utilizarse una regresión para evaluar efectos causales. Recordemos que en este ejemplo una diferencia de medias de la nota obtenida en el examen de selectividad por los individuos que reciben las clases de refuerzo y los que no las reciben no puede ser interpretada causalmente por la existencia de un sesgo de selección. A efectos ilustrativos, supongamos que las diferencias entre los alumnos que acuden a clases de refuerzo y los que no se deben únicamente a dos variables observables, la capacidad académica previa, y la renta de los padres (denotadas ambas por el vector X). Un modelo que tratase de medir el efecto del tratamiento (T) en el rendimiento académico (Y) podría representarse así:

$$Y = \beta_0 + \beta_1 T + \beta_2 X + e. \qquad [1]$$

En este modelo, e es un error que suponemos aleatorio y los coeficientes β_0, β_1, y β_2 son parámetros a estimar. El coeficiente de interés es β_1, que nos daría la diferencia entre haber recibido clases de refuerzo o no para individuos de igual renta o capa-

cidad. Es decir, el coeficiente β_1 capturaría el efecto causal del tratamiento, ya que se han controlado los efectos en la nota en el examen de selectividad que provienen de la distinta capacidad de cada estudiante y de la renta del hogar, las únicas variables que diferencian en nuestro ejemplo al grupo de control y al grupo de tratamiento.

Por el contrario, una regresión mínimo-cuadrática en la que únicamente se incluya como variable explicativa el tratamiento (haber recibido clases o no), puede representarse de la siguiente forma:

$$Y = \alpha + \beta T + e. \qquad [2]$$

Comparando la ecuación [1] con la [2], es fácil deducir que la estimación por MCO de la ecuación [2] daría lugar a la siguiente relación entre el parámetro estimado (β) y el verdadero efecto del tratamiento (β_1):

$$\beta = \beta_1 + \beta_2\, cov(T,X). \qquad [3]$$

En este caso, β sería un estimador sesgado, que incluiría el efecto verdadero del tratamiento y el de las otras variables que diferencian al grupo tratado y al de control. El sesgo vendría producido porque el valor de la variable tratamiento depende de la renta de los padres y de la capacidad académica de cada individuo; es decir, $cov(T,X)$ no sería igual a 0.

Resulta, en consecuencia, crucial asumir que el modelo incluye todas las variables de control que diferencian al grupo tratado y al grupo de control. Obviamente, esta hipótesis es difícil de cumplir en la realidad. Esto hace de los modelos de regresión una alternativa de estimación de efectos causales netamente inferior a la aleatorización.

2. Modelos *matching*

Un modelo *matching* o de emparejamiento consiste en emparejar a cada individuo del grupo de tratamiento, no control con otro individuo (pueden ser más de uno) que no ha recibido el tratamiento, pero que sea lo más parecido en su conjunto en términos de otras características observables distintas al tratamiento (Rubin, 1973; Heckman, Ichimura y Todd, 1998). Al igual que ocurre con los modelos de regresión, para que este tipo de modelos permita identificar el efecto causal del tratamiento es necesario que se cumpla la hipótesis de independencia condicional; es decir, que la asignación al grupo de tratamiento dependa únicamente de las variables observables que se utilizan para emparejar a los individuos. Sin embargo, al contrario de lo que sucede en los modelos de regresión, en los modelos de emparejamiento no es necesario asumir una determinada relación funcional entre las variables incluidas en el modelo, lo que reduce el riesgo de obtener resultados sesgados e inconsistentes.

Existen muchos métodos para realizar el emparejamiento entre individuos tratados y no tratados (Imbens, 2015; King y Nielsen, 2019). El método de emparejamiento más conocido es el *propensity score matchin* propuesto por Rosembaum y Rubin (1983). Este método utiliza un modelo de variable dicotómica (probit o logit) para calcular la probabilidad condicional de pertenecer al grupo de tratamiento dadas las variables observadas *X*. A la predicción resultante de ese modelo para cada individuo se le llama *propensity score*. A partir de ese *propensity score* se asocia a cada individuo del grupo tratado con aquel individuo del grupo de control que presente el valor del *propensity score* más cercano. El objetivo final es seleccionar una muestra de individuos no tratados que sea lo más parecida posible en sus características observables a la muestra de individuos tratados, algo que puede comprobarse fácilmente a través de distintos test de diagnóstico. Uno de los requisitos fundamentales que hay que comprobar tras realizar el emparejamiento es que se cumpla la condición del *common support:* debe haber individuos suficientemente parecidos en el grupo de control para cada individuo tratado.

En el caso español existen numerosas referencias a trabajos que utilizan los métodos de emparejamiento como técnica principal de análisis. Haremos referencia, a efectos ilustrativos, a dos de estos trabajos.

El primero de estos ejemplos, en el ámbito de las energías renovables, es el trabajo de Sánchez-Braza y Pablo-Romero (2014), que analiza las diferencias en la cantidad de metros cuadrados de tierra con placas solares entre los municipios que introdujeron en el año 2010 beneficios fiscales para aquellos que las instalaran frente a los municipios que no lo consideraron. Los autores encuentran un efecto positivo de la bonificación fiscal en la instalación de placas solares, que aumentaron en más de un 70 por 100 en los municipios tratados respecto a sus pares.

Otro ejemplo en el ámbito educativo es el trabajo de García-Pérez e Hidalgo-Hidalgo (2017), que evalúa la introducción del Programa de Acompañamiento Escolar, PAE, implantado de manera progresiva desde el año 2005 hasta el año 2012, sobre el rendimiento académico. Los estudiantes tratados eran aquellos individuos de las escuelas que participaban en el PAE el mismo año en que se hacían los exámenes del programa para la evaluación internacional de estudiantes (PISA por su denominación en inglés, Programme for International Student Assessment), principalmente en el año 2011/2012. Por el contrario, el grupo control estaba compuesto de aquellos estudiantes cuyos colegios no implantaron el PAE en absoluto, pero cuyas características observables eran similares según el modelo *matching* utilizado. Los autores concluyeron que pertenecer a dicho programa de acompañamiento reportó efectos positivos en el rendimiento académico, sobre todo para aquellos que llevaban más tiempo dentro de dicho programa.

En definitiva, los métodos *matching* son una técnica muy extendida en el análisis de las relaciones causa-efecto, ya que es relativamente sencilla de aplicar y permite obtener estimaciones insesgadas siempre y cuando se tenga información suficiente, al igual que en los modelos de regresión, sobre las variables observables que determinan la pertenencia al grupo de tratamiento.

VI. MÉTODOS CUASIEXPERIMENTALES

El principal problema tanto de las técnicas de emparejamiento como de los modelos de regresión es que solo permiten garantizar una estimación insesgada del efecto causal cuando se dispone de las variables observables necesarias. En muchos casos existen diferencias no observables entre el grupo de control y el grupo tratado que impiden que las estimaciones obtenidas con modelos de regresión o modelos de emparejamiento eliminen los sesgos de selección. En efecto, supongamos que queremos estimar el siguiente modelo econométrico:

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 N + e, \qquad [4]$$

donde Y, T y X tienen la misma interpretación que en la ecuación 1. Obviamente, como las características no observables, N, no pueden incluirse en el modelo, una estimación por MCO (o por métodos *matching*) que incluya únicamente las variables observables dará lugar a un sesgo de selección. En estos casos existen técnicas estadísticas que reciben el nombre de métodos cuasiexperimentales, que permiten eliminar los sesgos de selección, incluso cuando estos sesgos se deben a diferencias no observables entre el grupo tratado y el grupo de control. A continuación, describimos tres de estas técnicas.

1. Regresión discontinua

El diseño de regresión discontinua (RD) se ha utilizado en numerosos trabajos aplicados recientes, al ser uno de los métodos cuasiexperimentales más creíbles para la identificación, estimación e inferencia de los efectos de un tratamiento (Cook, 2008; Lee y Lemieux, 2010). Este tipo de diseño fue originariamente introducido por Thistlethwaite y Campbell (1960), quienes analizaron el efecto de los reconocimientos al mérito sobre los resultados académicos futuros, asumiendo que la asignación de estos reconocimientos estuviese basada en una nota de examen o prueba de evaluación observada. La hipótesis de los autores era que los individuos con notas justo por debajo del umbral y que no recibieron el premio servían como buenos sujetos con los que comparar aquellos cuya nota estaba justo por encima y efectivamente recibieron el reconocimiento. En realidad, en el entorno cercano al umbral de corte de la variable que determina el tratamiento (variable de asignación) puede suponerse que el tratamiento está aleatorizado de forma similar a como ocurre en un ensayo aleatorio.

Esta técnica puede utilizarse para estimar efectos causales «locales». Esto significa que únicamente alrededor del umbral de corte de la variable de asignación que determina el tratamiento, las diferencias en la variable objetivo que se encuentren pueden considerarse debidas al tratamiento, ya que para individuos alejados del umbral no puede asumirse que estar a un lado u otro del umbral sea aleatorio.

Para que la inferencia obtenida a través de un diseño de regresión discontinua sea válida es necesario mostrar que se cumplen una serie de condiciones (Lee y Lemieux, 2010; Abadie y Cattaneo, 2018), como son la ausencia de manipulación en la asignación a un grupo, la ausencia de diferencias en otras variables distintas a la variable objetivo en el entorno cercano al umbral, y la consistencia de los resultados tras la realización del test de «placebo».

Aun pudiendo estimarse tanto de forma paramétrica como con métodos no paramétricos, la forma más sencilla de estimar estos modelos es a través de una regresión del siguiente tipo restringida a observaciones cuyo valor en la variable de asignación esté comprendido dentro de un determinado rango alrededor del umbral:

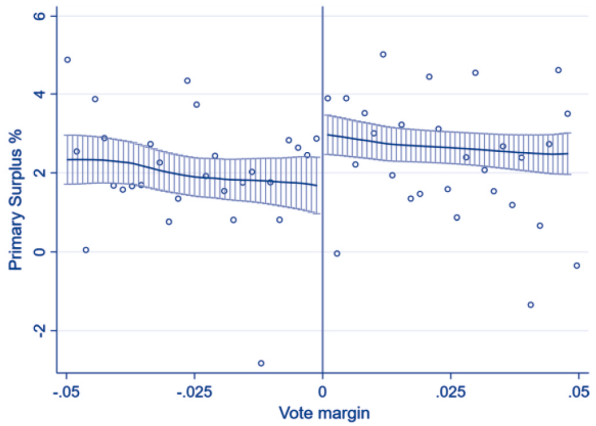
$$Y = \beta + \beta_1 T + \beta_2 \textit{ Variable de asignación } + e. \quad [5]$$

En esta regresión, al incluir como regresor la variable de asignación, el efecto del tratamiento, β_1, se estaría estimando de forma local; es decir, justo en el umbral que determina la pertenencia al grupo de tratamiento.

Si bien la estimación por MCO es muy sencilla, la forma más habitual de estimar un modelo de regresión discontinua es de forma no paramétrica; y los resultados se suelen representar gráficamente. Como ilustración, en el gráfico 1 mostramos los resultados de estimar un modelo no paramétrico de regresión discontinua extraído del trabajo de Artés y Jurado (2018). Este trabajo analiza el efecto en el gasto público y en los ingresos públicos de la existencia de Gobiernos de coalición frente a Gobiernos de un único partido en municipios españoles. La variable de «asignación» utilizada en este trabajo fue el porcentaje de votos que el partido ganador obtuvo por encima o por debajo del porcentaje mínimo de votos necesario para gobernar en solitario. El gráfico 1 ilustra el efecto causal de pasar de tener un Gobierno en coalición (lado izquierdo del umbral) a gobernar en solitario (lado derecho de la línea roja) sobre el excedente presupuestario (diferencia entre ingresos y gastos públicos). Puede observarse que justo en el umbral que determina el tratamiento (pasar de Gobierno de coalición a mayoría absoluta) se produce una discontinuidad o salto en el excedente presupuestario. La cuantificación de la magnitud y la significatividad del efecto en un modelo de regresión discontinua sería la diferencia entre la línea ajustada a un lado y a otro del umbral, pudiendo estimarse con un modelo no paramétrico. Una buena descripción del método puede encontrarse en Cattaneo *et al.* (2016).

Finalmente, concluimos este breve repaso a los diseños de regresión discontinua haciendo referencia a otro trabajo reciente que ha utilizado esta técnica en el caso español. Se trata del trabajo de Curto-Grau *et al.* (2018) que analiza las transferencias de gasto entre Gobiernos regionales y municipales. El trabajo muestra que los municipios ideológicamente alineados con el Gobierno regional obtienen transferencias por parte de ese Gobierno que son hasta un 100 por 100 superiores. Para identificar el efecto causal, los autores estudian únicamente Gobiernos municipales en los que el bloque ideológico que finalmente acabó gobernando se determinó por muy pocos votos, siendo la variable de asignación el número de votos que habrían hecho falta para que cambiase el bloque ideológico que controla el municipio. Los autores también muestran que el efecto es mucho mayor cuando el Gobierno regional gobierna con una mayoría holgada, llegando a desaparecer cuando el Gobierno regional obtuvo la mayoría por muy poco margen.

GRÁFICO 1
REPRESENTACIÓN GRÁFICA DE UN ESTUDIO EN EL QUE SE APLICA REGRESIÓN DISCONTINUA



Nota: El eje horizontal muestra el margen de votos respecto al número de votos para alcanzar la mayoría y gobernar en solitario, mientras que el eje vertical representa el excedente presupuestario.
Fuente: Artés y Jurado (2018).

2. Diferencias en diferencias

El método de diferencias en diferencias (*DiD*, por su denominación en inglés, *differences-in-differences*) es otra alternativa enormemente popular para evaluar el impacto causal de una política pública incluso en presencia de factores no observados. Esa metodología resulta especialmente adecuada cuando el tratamiento se aplica a un grupo de individuos, pero no a otros a partir de un determinado momento del tiempo (Heij *et al.*, 2004; Rabe-Hesketh y Skrondal, 2008). Normalmente, estas

políticas se aplican al grupo de tratamiento a partir de un momento del tiempo, por lo que se pueden observar las diferencias en la variable objetivo entre el grupo tratado y el grupo de no tratados antes de la implementación de la política pública, asumiendo que dichas diferencias debidas a variables no observables cuando no hay tratamiento son constantes a lo largo del tiempo.

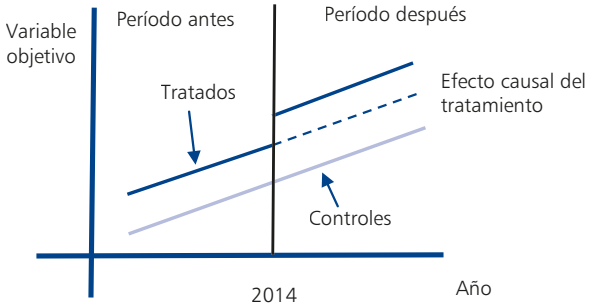
En términos de regresión, el modelo de diferencias en diferencias puede expresarse de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 \, Después + \beta_3 T_i * \\ * \, Después_i + \beta_4 X_i' + \varepsilon_i. \qquad [6]$$

Y_i representa la variable objetivo de que se trate. El vector X incluye variables de control que permiten controlar por diferencias observables entre cada individuo. T es una variable dicotómica que tiene valor 1 para los individuos que pertenecen al grupo de tratamiento y 0 para los que no. Nótese que la variable T tiene el mismo valor en el período antes del tratamiento y en el período posterior; es decir, no mide el efecto del tratamiento, sino que controla por las diferencias no observables que existen de forma permanente entre los individuos del grupo tratado y los demás. La variable *Después* es otra *dummy* que adopta valor 1 para los períodos posteriores a la implementación del tratamiento y 0 para los anteriores. Esta variable captura los *shocks* temporales que pudieran afectar a la variable objetivo a lo largo del tiempo y que son comunes tanto al grupo tratado como al grupo de control. La interacción de T y *Después* identifica a las observaciones del grupo tratado durante los períodos de tratamiento. β_3 es, en consecuencia, el coeficiente de interés, que permite aislar el efecto causal.

Al igual que ocurre con los diseños de regresión discontinua, para que la inferencia obtenida a través de un diseño de diferencias en diferencias sea válida es necesario mostrar que se cumplan una serie de condiciones. La más importante es la existencia de tendencias paralelas entre el grupo tratado y el grupo de control. Ambos grupos no tienen por qué ser iguales, de ahí la gran ventaja de este estimador, pero debe mostrarse que las diferencias en la evolución de la variable objetivo en el período pretratamiento son constantes a lo largo del tiempo. Existen varios test econométricos que permiten comprobar que se cumple esta condición, pero en muchas ocasiones es suficiente con un gráfico que muestre la evolución de la variable objetivo. En el gráfico 2 se ofrece una representación gráfica del modelo de diferencias en diferencias.

GRÁFICO 2
REPRESENTACIÓN GRÁFICA DEL MODELO DE DIFERENCIAS EN DIFERENCIAS



Fuente: Elaboración propia.

Supongamos una política pública que se aplica a un grupo tratado en el año 2014. Idealmente, el grupo de control (la línea azul claro) y el grupo tratado (línea azul oscuro) deben presentar una evolución similar en la variable objetivo antes del tratamiento. Si esto es así, el grupo de control sería un contrafactual adecuado del grupo tratado una vez que se tengan en cuenta las diferencias constantes entre ambos grupos a lo largo de todo el período.

En el contexto español existen numerosos trabajos que han utilizado la técnica de diferencias en diferencias para evaluar políticas públicas de los que, a modo de ejemplo, destacamos dos. En el reciente trabajo de García-Vega, Kneller y Stiebale (2021), los autores tenían como objetivo evaluar el efecto del contrato de emprendedores contemplado en la reforma laboral española del año 2012 en la innovación. Los autores concluyeron que el número de innovaciones de producto aumentó de manera significativa para las empresas tratadas (aquellas que se podían beneficiar de dicho contrato) en el período posterior a la reforma laboral introducida en España en el año 2012 respecto al grupo control.

Otro ejemplo es el análisis sobre el impuesto a las bebidas azucaradas implantado en marzo del año 2017 en Cataluña (Fichera *et al.*, 2021). Usando información sobre el consumo de 26 tipos de bebidas, tanto azucaradas como no azucaradas, los autores encuentran una reducción en el con-

sumo de bebidas azucaradas tras la introducción del tributo sobre este tipo de bebidas, mostrando diferencias según el tipo de bebida.

3. Variables instrumentales

Una tercera técnica que permite en principio estimar el efecto causal de un tratamiento en una variable de interés, incluso cuando existen variables no observables omitidas en el modelo, son los modelos de variable instrumental. Estos modelos se apoyan en la siguiente lógica: para comprobar si los cambios en una variable causan los cambios en la otra, hay que identificar una tercera variable (el instrumento) que afecte a la probabilidad de recibir el tratamiento, pero no a otras variables que puedan afectar a la variable de interés (Martens *et al.,* 2006; Imbens, 2014). La razón por la que esta variable debe afectar a la probabilidad de recibir el tratamiento, pero no a la variable de interés, es porque queremos comprobar, precisamente, si los cambios en el tratamiento provocados por cambios en esta tercera variable afectan a la variable de interés únicamente a través del efecto que ha producido el instrumento en el tratamiento. Si el instrumento afecta directamente tanto la variable de tratamiento como a la variable objetivo, el instrumento es inválido, ya que no permite distinguir si los cambios en la variable objetivo se deben a los cambios en el tratamiento o al efecto directo en la variable objetivo de esa tercera variable. Si el instrumento solo afecta directamente al tratamiento, entonces sí observamos un cambio en la variable objetivo cuando cambia el instrumento y, por tanto, podemos estar seguros de que la causa del cambio en la variable objetivo es precisamente el tratamiento.

Una vez aclarado intuitivamente en qué consiste un modelo de variable instrumental, debe aclararse también que una variable solo puede servir como instrumento cuando presenta dos características estadísticas: en primer lugar, la validez (ha de afectar solo al tratamiento, pero no directamente a la variable objetivo) y, en segundo lugar, la fortaleza (debe afectar lo suficiente a la probabilidad de recibir el tratamiento como para poder extraer conclusiones estadísticamente relevantes).

Un ejemplo clásico del uso de variable instrumental es el análisis de Angrist y Kruger (1991) del efecto de los años de escolarización en el salario de los individuos. Angrist y Krueger observaron que la fecha de nacimiento de cada niño determina exógenamente los años de escolarización. En efecto, pensemos en el caso de España: un niño nacido en enero recibirá, en media, un año más de escolarización obligatoria que un niño nacido en diciembre. Los autores argumentan que como la fecha de nacimiento no tiene efectos en la habilidad o capacidad innata del individuo, característica no observable que codetermina ambas variables, puede utilizarse como instrumento ya que cumpliría las condiciones de validez y fortaleza. El trabajo muestra, utilizando datos de EE. UU., que los años adicionales de escolarización tienen un efecto positivo y significativo en el salario posterior de los individuos.

Un ejemplo reciente de utilización de variable instrumental para el caso español es Artés (2014). En este trabajo se trata de analizar a qué partido político favorece la abstención estimando un modelo en el que la variable dependiente es el porcentaje de voto a distintos partidos en cada municipio en distintas elecciones generales. La variable de interés o de tratamiento es el porcentaje de abstención. El instrumento utilizado es la meteorología; en concreto, el porcentaje de lluvia en cada municipio en el día de las elecciones. Este instrumento afecta a la abstención (los días de lluvia es más costoso ir a votar, por lo que existe una relación clara entre la lluvia y el porcentaje de participación electoral), pero no al sentido del voto (las personas que acuden a votar no deciden a quién votar en función de si llueve o no ese día). El cumplimiento de la condición de fortaleza y validez permite estimar causalmente el efecto de la abstención en el porcentaje de voto a cada partido. El autor encuentra que la abstención, en general, beneficia a los partidos de derechas, y perjudica, en mayor medida, a los partidos no tradicionales.

VII. DESARROLLOS RECIENTES

En este trabajo se ha mostrado no solo que las técnicas para evaluar los posibles efectos de las políticas públicas son numerosas, sino que experimentan una constante evolución y desarrollo. En línea con algunos de los métodos antes descritos, en esta sección se van a mencionar brevemente algunos desarrollos metodológicos recientes en este ámbito, que están en la mayoría de los casos relacionados de una manera u otra con las técnicas antes descritas.

En el ámbito de los modelos *matching*, existen una variedad de desarrollos recientes que tratan de mejorar algunas limitaciones de los modelos tradicionales. Una de estas técnicas es el emparejamiento

genético o *genetic matching* que corrige algunas limitaciones de los modelos que utilizan el *propensity score* (Diamond y Sekhon, 2013). En primer lugar, dicha técnica elimina la necesidad de verificar de forma manual e iterativa el *propensity score*, ya que, por un lado, proporciona un algoritmo de búsqueda automatizado que identifica las mejores coincidencias y maximiza el equilibrio en las variables observables entre el grupo tratado y el grupo de control. Esto supone estimaciones menos sesgadas (Radice *et al*., 2011; Diamond y Sekhon, 2013).

Siguiendo una lógica similar a los modelos de diferencias en diferencias, el método del control sintético ha sido descrito como «el desarrollo más importante en la evaluación de programas en la última década» (Athey e Imbens, 2017). Esta técnica, a grandes rasgos, consiste en construir un grupo de control «sintético» asignando distintos pesos a las observaciones no tratadas (a diferencia de modelo de diferencias en diferencias en el que cada observación tiene el mismo peso) de forma que en su conjunto se consiga un contrafactual lo más similar posible al grupo tratado. Una descripción de las ventajas de este método y de sus aplicaciones puede encontrarse en Abadie (2021). Otro desarrollo reciente en el ámbito de los estimadores de diferencias en diferencias son las mejoras en la estimación de modelos en los que el tratamiento se aplica a distintas unidades del grupo tratado en distintos momentos del tiempo. En estos casos la hipótesis de que las tendencias son paralelas no siempre permite identificar el efecto causal del tratamiento, por lo que se han desarrollado métodos que permiten superar estos problemas como los desarrollados por Sant'Anna y Zhao (2020), Callway y Sant'Anna (2021) o Athey e Imbens. (2022)

Finalmente, cabe mencionar la utilidad para los análisis causa-efecto de desarrollos recientes en el ámbito de los modelos de *big data*. Las técnicas de *machine learning* consisten en utilizar una submuestra, llamada de «entrenamiento» o *training sample*, para predecir el comportamiento de una variable de interés. Una vez que se ha validado la capacidad predictiva en la submuestra, puede aplicarse ese modelo para predecir la evolución de la variable objetivo a gran escala. Aunque estas técnicas tienen como objetivo la predicción y no la identificación de efectos causales, tienen un potencial enorme para ayudar a construir grupos de control enormemente parecidos al grupo tratado que se quiere analizar, permitiendo que el contrafactual estimado se acerque al ideal de la aleatorización en mayor medida. Una revisión de estos métodos y de su potencial para mejorar la evaluación de políticas públicas puede encontrarse en Athey e Imbens (2017: pp. 22-27).

VIII. CONCLUSIONES

La evaluación del impacto causal de las políticas públicas es fundamental para mejorar su implementación. En este trabajo se ha hecho una revisión de los principales métodos de estimación de relaciones causa-efecto en este ámbito. Todos estos métodos pretenden encontrar un contrafactual adecuado que permita estimar qué habría ocurrido con la variable de interés si no se hubiera llevado a cabo la política pública analizada.

La aleatorización *ex ante* del grupo tratado y del grupo de control es el mejor método para encontrar un contrafactual adecuado. Sin embargo, como la aleatorización no es posible en la gran mayoría de casos, se han desarrollado técnicas que bajo determinadas condiciones permiten la identificación precisa de los efectos causales de una política pública, incluso cuando existen diferencias no observables entre el grupo de individuos beneficiario de la política pública y los no beneficiarios.

En la actualidad, la utilización de estos métodos en España no es muy frecuente más allá del ámbito académico. Sería deseable su aplicación en el ámbito de la gestión pública, tanto para mejorar el diseño como la aplicación de las políticas públicas. Un conocimiento adecuado de los efectos de los programas públicos redundaría en una mejora de la toma de decisiones por parte de los gestores políticos y en consecuencia del bienestar general.

NOTAS

(*) Joaquín Artés agradece la financiación recibida a través del proyecto CSO2017-82881_R.

(1) En el año 2019 se le concedió el premio Nobel de Economía a Esther Duflo, Abhijit Banerjee y Michael Kremer por el uso de ensayos aleatorizados (*randomized controlled trials,* o *RCT*, por sus siglas en inglés) en el ámbito de las políticas públicas.

(2) Una descripción del programa y una lista completa de publicaciones relacionadas puede consultarse en la página del NBER dedicada a este experimento: https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment?page=1&perPage=50

BIBLIOGRAFÍA

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2), pp. 391-425.

ABADIE, A. y CATTANEO, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10, pp. 465-503.

ALBERT, R., ESCOT, L. y FERNÁNDEZ-CORNEJO, J. A. (2011). A field experiment to study sex and age discrimination in the Madrid labour market. *The International Journal of Human Resource Management*, 22(2), pp. 351-375.

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.

ARTÉS, J. (2014). The rain in Spain: Turnout and partisan voting in Spanish elections. *European Journal of Political Economy*, 34, pp. 126-141.

ARTÉS, J. y JURADO, I. (2018). Government fragmentation and fiscal deficits: a regression discontinuity approach. *Public Choice*, 175(3), pp. 367-391.

ATHEY, S. e IMBENS, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), pp. 3-32.

ATHEY, S. e IMBENS, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics,* 226(1), pp. 62-79.

BAGUES, M. F. y ESTEVE-VOLART, B. (2010). Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *The Review of Economic Studies*, 77(4), pp. 1301-1328.

BERTRAND, M. y MULLAINATHAN, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), pp. 991-1013.

CALLAWAY, B. y SANT'ANNA, P. HC. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225, pp. 200-300.

CATTANEO, M. D., TITIUNIK, R., VÁZQUEZ-BARE, G. y KEELE, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4), pp. 1229-1248.

COOK, T. D. (2008). «Waiting for life to arrive»: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2), pp. 636-654.

CURTO-GRAU, M., SOLÉ-OLLÉ, A. y SORRIBAS-NAVARRO, P. (2018). Does electoral competition curb party favoritism? *American Economic Journal: Applied Economics*, 10(4), pp. 378-407.

DIAMOND, A. y SEKHON, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), pp. 932-945.

DÍAZ-SERRANO, L. y MEIX-LLOP, E. (2016). Do schools discriminate against homosexual parents? Evidence from a randomized correspondence experiment. *Economics of Education Review*, 53, pp. 133-142.

FICHERA, E., MORA, T., LÓPEZ-VALCÁRCEL, B. G. y ROCHE, D. (2021). How do consumers respond to «sin taxes»? New evidence from a tax on sugary drinks. *Social Science & Medicine*, 274, 113799.

FINKELSTEIN, A., TAUBMAN, S., WRIGHT, B., BERNSTEIN, M., GRUBER, J., NEWHOUSE, J. P., ALLEN, H., BAICKER, K. y OREGON HEALTH STUDY GROUP (2012). The Oregon health insurance experiment: evidence from the first year. *The Quarterly Journal of Economics*, 127(3), pp. 1057-1106.

GARCÍA-PÉREZ, J. I. e HIDALGO-HIDALGO, M. (2017). No student left behind? Evidence from the Programme for School Guidance in Spain. *Economics of Education Review*, 60, pp. 97-111.

GARCÍA-VEGA, M., KNELLER, R. y STIEBALE, J. (2021). Labor Market reform and innovation: Evidence from Spain. *Research Policy*, 50(5), 104213.

HECKMAN, J. J., ICHIMURA, H. y TODD, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2), pp. 261-294.

HEIJ, C., DE BOER, P., FRANSES, P. H., KLOEK, T. y VAN DIJK, H. K. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

IMBENS, G. W. (2014). Instrumental variables: an econometrician's perspective (n.º w19983). *National Bureau of Economic Research*.

IMBENS, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2), pp. 373-419.

KING, G. y NIELSEN, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4), pp. 435-454.

KRUEGER, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2), pp. 497-532.

LEE, D. S. y LEMIEUX, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), pp. 281-355.

MARTENS, E. P., PESTMAN, W. R., DE BOER, A., BELITSER, S. V. y KLUNGEL, O. H. (2006). Instrumental variables: application and limitations. *Epidemiology*, 260-267.

PEÑA-LONGOBARDO, L. M., OLIVA-MORENO, J., ZOZAYA, N., ARANDA-RENEO, I., TRAPERO-BERTRAN, M., LAOSA, O., SINCLAIR, A. y RODRÍGUEZ-MAÑAS, L. (2021). Economic evaluation of a multimodal intervention in pre-frail and frail older people

with diabetes mellitus: the MID-FRAIL project. *Expert Review of Pharmacoeconomics & Outcomes Research*, 21(1), pp. 111-118.

Rabe-Hesketh, S. y Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., y Sekhon, J. S. (2012). Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *The International Journal of Biostatistics*, 8(1), p. 25.

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

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1), pp. 185-203.

Sánchez-Braza, A. y Pablo-Romero, M. D. P. (2014). Evaluation of property tax bonus to promote solar thermal systems in Andalusia (Spain). *Energy Policy*, 67, pp. 832-843.

Sant'Anna, P. H. y Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics,* 219(1), pp. 101-122.

Thistlethwaite, D. L. y Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), p. 309.