Sobre el uso de la evidencia y la validez externa en la evaluación de intervenciones sociales: una mirada crítica

Juan David Parra
Universidad del Norte (Colombia)

CÓMO CITAR:
Parra, Juan David. 2021. "Sobre el uso de la evidencia y la validez externa en la evaluación de intervenciones sociales: una mirada crítica". *Colombia Internacional* 105: 175-198. https://doi.org/10.7440/colombiaint105.2021.07

RECIBIDO: 26 de febrero de 2019
ACEPTADO: 22 de noviembre de 2019
MODIFICADO: 28 de noviembre de 2019
https://doi.org/10.7440/colombiaint105.2021.07

RESUMEN: **Objetivo/contexto**: son cada vez más las voces críticas sobre la idoneidad de prácticas dominantes en la evaluación de intervenciones sociales (por ejemplo, las RCT) frente al objetivo de informar políticas basadas en evidencia. Este artículo se enfoca en dos elementos centrales de dicha reflexión: i) la evidencia, como concepto y como resultado de un proceso de razonamiento; y ii) la noción de validez externa. **Metodología**: el uso de literatura en los campos de la evaluación y de la filosofía del conocimiento me permite hacer una deconstrucción del concepto de *causalidad* en las ciencias sociales. A partir de este ejercicio, y de la distinción entre teorías de causalidad *sucesionista* y *generativa*, identifico criterios para examinar críticamente algunos postulados epistemológicos implícitos en argumentos de exponentes de las técnicas experimentales de evaluación. **Conclusiones**: las técnicas experimentales no permiten, por sí mismas, informar decisiones sobre cómo invertir recursos de forma eficiente. Pese a su fortaleza para cuantificar posibles efectos causales, es necesario complementar el análisis estadístico contrafactual con formas de razonamiento cualitativo conducentes a solucionar interrogantes sobre las causas eficaces detrás del resultado de intervenciones sociales, y los factores de soporte que podrían permitir pensar en la extrapolación de políticas o programas sociales entre diferentes contextos. **Originalidad**: la literatura en español sobre críticas y alternativas a las técnicas de evaluación de impacto es escasa. Más que presentar un resumen de argumentos de otros autores, este artículo construye una narrativa coherente para repensar el papel de la evaluación en la sociedad.

PALABRAS CLAVE: evaluación; técnicas experimentales; evidencia; validez externa.

On the Use of Evidence and External Validity in the Evaluation of Social Interventions: A Critical Overview

ABSTRACT: **Objective/Context:** There is a growing criticism of the mainstream evaluations of the social interventions (e.g. RCTs) which shape evidence-based policies. This article focuses on two central aspects of this question: i) the notion

of evidence, as a concept and a result of a process that entails reasoning, and ii) the understanding of external validity. **Methodology:** Through an analysis of the literature on evaluation and the philosophy of knowledge, I deconstruct the concepts of *causality* in the social sciences. This exercise allows me to distinguish between *successionist* and *generative* theories of causality and establish criteria to critically examine some epistemological postulates of experimental evaluation techniques. **Conclusions:** By themselves, experimental methods of evaluation do not allow for informed decisions about an efficient investment of resources. Despite their strength in quantifying possible causal effects, counterfactual statistical analyses need to be complemented by forms of qualitative reasoning in order to answer questions about the direct and indirect causes of the results of social interventions, and thus strengthen our understanding of the extrapolation of social policies or programs from certain contexts. **Contribution:** There are few studies in Spanish which criticize experimental evaluation techniques or suggest alternatives to them. Instead of summarizing the arguments of other authors, this article presents a coherent narrative which asks us to rethink the current role of the evaluation of social interventions.

KEYWORDS: Evaluation; experimental techniques; evidence; external validity.

Sobre o uso de evidências e validade externa na avaliação das intervenções sociais: um olhar crítico

RESUMO: **Objetivo/contexto:** há cada vez mais vozes críticas sobre a adequação das práticas dominantes na avaliação das intervenções sociais (por exemplo, RCT) versus o objetivo de informar políticas baseadas em evidências. Este artigo enfoca dois elementos centrais de tal reflexão: i) a evidência, como um conceito e como resultado de um processo de raciocínio e ii) a noção de validade externa. **Metodologia:** o uso da literatura nas áreas de avaliação e filosofia do conhecimento me permite fazer uma desconstrução do conceito de *causalidade* nas ciências sociais. A partir desse exercício e da distinção entre teorias de causalidade *sucessória* e *generativa*, identifico critérios para examinar criticamente alguns postulados epistemológicos implícitos em argumentos de expoentes de técnicas de avaliação experimental. **Conclusões:** as técnicas experimentais, por si só, não informam as decisões sobre como investir recursos de forma eficiente. Apesar de sua força em quantificar possíveis efeitos causais, é necessário complementar a análise estatística contrafactual com formas de raciocínio qualitativo que conduzam à solução de questões sobre as causas efetivas por trás do resultado das intervenções sociais e os fatores de apoio que tornem possível pensar sobre a extrapolação de políticas ou programas sociais entre diferentes contextos. **Originalidade:** a literatura, principalmente em espanhol, sobre críticas e alternativas às técnicas de avaliação de impacto é escassa. Em vez de apresentar um resumo dos argumentos de outros autores, este artigo constrói uma narrativa coerente para repensar o papel da avaliação na sociedade.

PALAVRAS-CHAVE: avaliação; técnicas experimentais; evidências; validade externa.

Introducción

Este artículo responde a un interrogante implícito en todo ejercicio de evaluación de intervenciones sociales: ¿cómo usar los hallazgos para determinar qué hacer con un programa o proyecto exitoso? La reflexión puede también plantearse en términos de la ausencia de identificación de un efecto en un contexto determinado. Es decir, dado que una intervención X no arrojó resultados en un escenario Z, ¿debe descartarse que pueda tener impactos favorables en un escenario W? Estas son preguntas importantes para una sociedad en general, y para una oficina o institución ejecutora de recursos públicos en particular, a fin de dar cumplimiento al anhelo de la llamada política basada en evidencia (PBE) de "verificar y mejorar la calidad, la eficiencia y la efectividad de las intervenciones [por ejemplo gubernamentales] en varias etapas de la implementación" (Gertler *et al.* 2011, 4). Sin embargo, este es un debate por lo general omitido, o en el mejor de los casos pormenorizado, entre académicos y practicantes expertos que acogen con entusiasmo el despliegue de *pruebas controladas aleatorizadas* (RCT, por sus siglas en inglés). Cobra pertinencia, por tanto, la advertencia que hace un premio Nobel de Economía:

> las RCT […] incluso aquellas hechas sin errores o contaminación, tienen una baja probabilidad de ser útiles para la política [pública], o para ir más allá de lo local, a menos que nos digan algo sobre por qué el programa funcionó, algo para lo cual, por lo general, no se encuentran orientadas o equipadas […] [La] validez de la política basada en la evidencia depende del eslabón más débil de la cadena de argumentos y de evidencia, de modo que para el momento que buscamos usar resultados experimentales, las ventajas de las RCT sobre […] otros métodos econométricos se habrán evaporado. Al final, no existe un sustituto a una evaluación cuidadosa de la cadena de evidencia y del razonamiento de personas con experiencia y pericia en el área [de estudio]. La necesidad de que los experimentos sean orientados por una teoría no representa, claramente, una garantía de éxito, pero la falta de ello es cercana a un garante de fracaso.[1] (Deaton 2010, 448-450)

Esta cita trae a colación varios elementos relevantes para la discusión. Asumamos de momento que una RCT se encuentra técnicamente bien diseñada, permitiendo así confiar en la validez interna de la evaluación. Lo anterior implica reconocer la fortaleza de los métodos experimentales para cuantificar y

1 Todas las traducciones son propias.

dimensionar efectos de una intervención en un contexto específico. Sin embargo, para Deaton (2010), la solidez estadística (por ejemplo, la eliminación de sesgos de estimación o la representatividad de los resultados) no garantiza que los hallazgos de lo que funciona en un contexto Z sean extrapolables a otros contextos ni tampoco que una intervención pueda seguir impactando positivamente una misma población. Parafraseando a Cartwright (2012, 2013), existe, por tanto, una brecha entre afirmaciones del tipo *el programa funciona en algún lugar* y *el programa funcionará acá*, que se explica porque los experimentos sociales no permiten entender *el porqué* y *el cómo* una política específica impacta, o no, a un grupo de personas. No es de extrañar, por tanto, que a pesar de los esfuerzos de gobiernos como, por ejemplo, el mexicano, por institucionalizar la evaluación al servicio del mejoramiento de las políticas públicas, "no [haya] evidencias contundentes de que las evaluaciones […] estén contribuyendo a incrementar la eficiencia de los programas sociales" (Cardozo 2013, 123). La revisión de literatura internacional de Head (2016) y el reciente comentario de Krause y Hernández (2020) plantean una conclusión similar.

Para algunos expertos, dicho debate implica regresar a preguntas básicas sobre el propósito de hacer una evaluación o sobre la esencia del análisis causa-efecto en ámbitos humanos. Trabajos como el de Pawson y Tilley (1997) y Cartwright y Hardie (2012), entre otros, ofrecen una base conceptual para abordar dicha reflexión. El análisis de este artículo se desprende de dos elementos concretos discutidos por estos autores. El primero tiene que ver con la importancia de resaltar el papel de los seres humanos como la *causa eficaz* del resultado de todo proyecto o política social. Según los críticos de las RCT, al centrarse solamente en las condiciones necesarias, y no en las motivaciones y los cursos de acción efectivamente tomados por individuos y grupos sociales, los esquemas de evaluación dominante dejan a un lado un componente esencial en el estudio de las intervenciones sociales (Pawson 2013). El segundo elemento hace referencia a un supuesto implícito del paradigma de la PBE sobre el carácter *automático* y *transferible* de los hallazgos de una evaluación para informar decisiones en torno al uso eficiente de recursos (Greenhalgh y Russell 2009). Sin embargo, la *validez externa* (como se reconoce en el lenguaje econométrico) no puede reducirse a un argumento fundamentado en la simple representatividad estadística de un estimador de impacto (Cartwright 2017). Sin hipótesis concretas que permitan identificar *evidencia creíble* sobre los mecanismos que explican por qué falla o no una intervención un escenario Z, y sin la existencia de factores de soporte que permitan especular que esta podría funcionar también en un escenario W, se hace (aún más) difícil predecir qué sucederá al escalar una intervención de un contexto a otro (Joyce y Cartwright 2020).

La contribución de este artículo consiste en presentar una serie de elementos heurísticos útiles para evaluadores profesionales y que sean conducentes a la generación de recomendaciones de política más factibles de ser implementadas con éxito. No hablamos en términos de minimizar el riesgo, ni tampoco de garantizar una buena predicción, en tanto esas son expectativas irreales en el estudio de sociedades abiertas y cambiantes (Parra 2016, 2018). Sin embargo, "es mejor estar más o menos en lo correcto, que precisamente equivocados" (Lawson 2009, 405), aforismo que indica que si bien el mejorar el uso que le damos a la evidencia no asegura que vayamos a tener siempre políticas exitosas, podemos eludir, al menos, rutas directas al fracaso (Reiss 2018).

1. De condiciones necesarias a poderes causales eficaces

Como se anticipó en la introducción, una reflexión sobre la relevancia de las recomendaciones que surgen de ejercicios de evaluación invita a examinar un concepto básico: la causalidad. Sin necesidad de traer a colación elementos del debate más amplio sobre la posibilidad de identificar relaciones causa-efecto en las ciencias sociales,[2] podemos sacar provecho de la distinción que hacen Pawson y Tilley (1997) entre una visión *sucesionista* y una *generativa* de la causalidad en la teoría de la evaluación. Según estos autores,

> [e]stas dos reconstrucciones resumen mucho de lo que está en disputa en nuestra comprensión de la causalidad […] Los sucesionistas siguen a Hume (1739) en su opinión de que la causalidad es inobservable (una *percepción de la mente*) y que solo se pueden hacer tales inferencias sobre la base de datos de observación. La clave es establecer una secuencia controlada de observaciones que diferencie la relación causal de la asociación espuria […] [La] otra gran teoría metafísica de la causalidad se remonta a la antigüedad filosófica, pero ha sido testigo de un desarrollo más sostenido en las últimas tres décadas. También enfatiza la necesidad de observar patrones regulares entre entradas y salidas, entre causas y sus efectos, pero busca establecer la conexión de una manera bastante diferente. La teoría generativa sostiene que hay una conexión real entre los eventos que entendemos que están conectados causalmente. […] [Por tanto], la teoría generativa considera que la causalidad actúa tanto interna como externamente [sobre el objeto o la política investigada]. La causa describe el potencial transformador de los fenómenos. Un hecho

2 Para profundizar en este debate, ver Parra (2016).

> bien puede desencadenar otro, pero solo bajo las condiciones adecuadas y en las circunstancias correctas. A menos que la explicación penetre en estos niveles subyacentes reales, se considera que está incompleta. (32-34; énfasis en el original)

Esta descripción básica permite concluir que las RTC (por sí mismas) reproducen la lógica de la causalidad sucesionista en la forma de A→B. Así lo presentan importantes exponentes de las denominadas técnicas de evaluación de impacto, para quienes los métodos experimentales aleatorios representan la mejor práctica posible para determinar si una intervención o programa tuvo un impacto causal sobre un grupo poblacional específico. En términos planos, "el impacto causal (α) de un programa (P) sobre un resultado (Y) es la diferencia entre el resultado (Y) con el programa (es decir, cuando P = 1) y el mismo resultado (Y) sin el programa (es decir, cuando P = 0)" (Gertler *et al.* 2011, 34). Dicho razonamiento no hace referencia al potencial transformador de los seres humanos afectados por el programa evaluado ni tampoco indaga sobre procesos o condiciones que permitan reflexionar en torno a cómo o por qué α puede arrojar un signo positivo y estadísticamente significativo. Para ser justos, los autores de esta última cita lo reconocen. Para Gertler *et al.* (2011),"[s]i bien las evaluaciones de impacto pueden producir estimaciones confiables de los efectos causales de un programa, su diseño no está normalmente orientado a informar la implementación del programa" (15). Por tanto, agregan, sin ser muy específicos al respecto, que "el trabajo cualitativo *puede contribuir* a que los responsables de políticas comprendan lo que está ocurriendo en el programa" (17; énfasis agregado).

Cabe recordar que este esquema experimentalista de evaluación tuvo su origen en estudios de medicina de mediados del siglo XX, y solo en décadas recientes se ha convertido en un referente de práctica ideal (una especie de *patrón-oro*) en el análisis de intervenciones sociales. Pese a ello, y de la misma forma en que investigadores en ciencias médicas han empezado a cuestionar seriamente los beneficios de las RCT como medio para acumular conocimiento en el uso de medicamentos y tratamientos clínicos (Greenhalgh y Russell 2009; Krauss 2018; Van Belle *et al.* 2016), las críticas a la posibilidad de aprender del mundo social a partir del análisis cuantitativo contrafactual son cada vez más agudas (Deaton 2010; Deaton y Cartwright 2018; Goodman, Epstein y Sullivan 2018, Joyce y Cartwright 2020; Saltelli y Giampietro 2017). Sugiere Pawson (2013) que parte del problema reside en que lo que la lógica experimental concibe como algo único y estable es, en realidad, parte de un proceso largo y continuamente cambiante de implementación. Por tanto, la confianza que tienen los evaluadores

experimentales en la uniformidad del mundo carece de sustento. Con ello coinciden personalidades como Hausmann (2016), de la Universidad de Harvard:

> [el] problema más importante que yo tengo con los RCT es que nos hacen pensar sobre las intervenciones, las políticas y las organizaciones de manera errónea. A diferencia de los dos o tres diseños que se prueban lentamente a través de un RCT (como colocar tabletas o papelógrafos en las escuelas), la mayor parte de las intervenciones sociales *tiene millones de posibilidades de diseño* y los *resultados dependen de combinaciones complejas entre ellas*. (Énfasis agregado)

La visión generativa de la causalidad representa, por tanto, una aproximación (de momento conceptual) alternativa a los problemas inherentes a lógicas sucesionistas de la causalidad al servicio de la evaluación de intervenciones sociales. Es relevante anotar que, pese a la defensa de exponentes de las técnicas experimentales frente al soporte científico de sus métodos de investigación —en medio del cual apelan a nociones de rigor numérico como garante de esa cientificidad—,[3] el proceder de estrategias como las RCT dista de emular las lógicas reales bajo las cuales opera el procedimiento de descubrimiento científico en disciplinas como la física o la química (Parra 2016). De hecho, según escribe Ellis (2005) en la revista *Nature*, los físicos contemporáneos se cuestionan sobre asuntos similares. A juicio del investigador, la visión cartesiana (o lineal) que impacta profundamente el estudio de átomos y partículas es insuficiente para escalar el análisis al estudio de organismos más complejos (por ejemplo, los seres humanos). La visión del experimento de laboratorio omite aspectos importantes del mundo con los que la física aún debe conciliarse, como por ejemplo el hecho de que nuestro ambiente se encuentre dominado por *cosas* que materializan los resultados de un diseño *intencionado* (un edificio, un libro, un computador, una cuchara). Ellis (2005) dictamina, por tanto, que "la física de hoy no tiene nada que decir frente a dicha intencionalidad que resulta en la existencia de tales objetos, a pesar de que dicha intención es, claramente, causalmente efectiva" (743).[4]

Una diferencia crucial entre ambas teorías de la causalidad (sucesionista y generativa) reside, por tanto, en el papel que atribuye el evaluador tanto al

---

3 Según sostienen Gertler *et al.* (2011), "[l]os analistas de datos son expertos en estadísticas y econometría que utilizan *software* de estadística como Stata, SPSS o R para analizar la evaluación de impacto. Los analistas de datos son responsables de *garantizar la calidad, el rigor científico y la credibilidad de los resultados*" (212-213; énfasis agregado).

4 Harman (2018) hace una reflexión similar sobre los límites de la física al momento de elegir entre alternativas teóricas en el campo de la *teoría de las cuerdas*.

contexto como a la intención y conducta individual o colectiva en el estudio de un programa o intervención. Lo anterior no es trivial. Similar al caso de un barril de pólvora, el cual "no siempre se enciende cuando se le aplica una llama", menos si "la mezcla está húmeda, si no hay suficiente [compuesto químico], si la mezcla no es compacta, si no hay oxígeno, si el calor se le aplica durante un periodo de tiempo insuficiente, etc." (Manzano 2010, 25), una política o programa social no va a tener efecto a menos que se creen motivaciones concretas a las poblaciones receptoras para que efectivamente hagan uso de los beneficios que les son ofrecidos. Es importante distinguir, por tanto, entre las condiciones necesarias que generan incentivos para cierto tipo de conducta y el curso de acción concreto (o los *poderes causales eficaces*) que surge a través del proceso. Al momento de una evaluación, es importante analizar ambos elementos y su interacción; si bien existen factores de contexto que restringen o premian diferentes acciones grupales o individuales (como la disponibilidad de recursos o las reglas de juego), los recursos de una intervención solo se transforman en resultados (observables, medibles) como resultado de la acción humana intencionada (Pawson 2013; Pawson y Tilley 1997).

Una debilidad concreta de la lógica sucesionista de herramientas como las RCT es que ignora por completo tal distinción (e interacción). Como lo advierten Porter, McConnell y Reidcor (2017), al reducir la explicación de una política o programa a su diseño —es decir, a los incentivos que *se presume* que este debe generar—, las técnicas experimentales de evaluación ignoran por completo la posibilidad de acción humana intencionada.[5] Un examen crítico del uso del concepto de la *teoría de cambio* (TC),[6] considerado por evaluadores de la tradición experimental como una herramienta esencial para guiar "la recolección y la interpretación de los resultados" (Bernal y Peña 2011, 7), permite ilustrar este último punto:

> El enfoque de la TC […] es [poco] indicativo sobre cómo debe acumularse el conocimiento dentro y entre las evaluaciones. Sin embargo, existe una suposición implícita de que esas teorías que se rastrean prospectivamente y se consideran especímenes resistentes se toman como dadas en evaluaciones futuras, liberando así al evaluador de esas teorías sobre las que se

5 De igual forma, según resaltan estos últimos autores, el énfasis que se da al comportamiento promedio (estadísticamente hablando) muestra que dichos métodos simplemente no se encuentran equipados para estudiar la voluntad individual.

6 En términos simples, los manuales de las técnicas de evaluación de impacto definen una TC como "la lógica causal de cómo y por qué un proyecto, un programa o una política lograrán los resultados deseados o previstos" (Gertler *et al.* 2011, 22). De la discusión del texto se deduce que esta es una definición inmersa en una noción sucesionista de la causalidad.

> sabe poco […] Un enfoque de TC, argumentamos, se preocupa más por los resultados generales del programa y las sinergias entre los diversos aspectos de una intervención. Por lo tanto, ayuda a proporcionar una perspectiva estratégica en un programa complejo […] Sin embargo, la desventaja es que las teorías descubiertas son relativamente superficiales y es más probable que permanezcan en el nivel de implementación. (Blamey y Mackenzie 2007, 448, 450-451)

Dicho énfasis en el deber ser de la implementación de una intervención (por ejemplo, si existen buenos docentes, los estudiantes se van a esforzar más y van a obtener mejores resultados en su exámenes) ha sido descrito por Deaton y Cartwright (2018) como una adherencia, al menos implícita,[7] a modelos simples del actor racional en la teoría económica (neoclásica) del consumidor.[8] La siguiente sección enfatiza en por qué esta condición ontológica de las técnicas de evaluación dominantes —según la cual cuentan con una limitación constitutiva para explicar el cómo y el porqué de los posibles resultados de una intervención social— les permite a sus promotores circunscribirse a una visión de la evidencia, en el mejor de los casos, acotada. A partir de ello, la tercera sección del artículo utiliza elementos de dicho debate para problematizar la noción de validez externa —entendida como la extrapolación de tipos de evidencia—.

2. La noción de evidencia como la búsqueda de poderes causales eficaces

El mensaje de la sección anterior es claro: sin un esfuerzo explícito por entender las motivaciones reales y las acciones concretas de los beneficiarios de una intervención social, perdemos de vista elementos básicos para explicar el cómo y el porqué detrás de sus resultados. El enfoque sucesionista de las RCT, según se discutió arriba, no es conducente a resolver ese tipo de interrogantes. A pesar de ello, sugieren algunos de sus exponentes, "[l]as evaluaciones bien diseñadas y bien implementadas pueden aportar *evidencias convincentes* y exhaustivas útiles para informar […] decisiones sobre políticas" (Gertler *et al*. 2011, 5-6; énfasis

7 No es explícita, en la medida en que, al menos según algunos de los exponentes de las técnicas experimentales, estas tienen la ventaja de no requerir teorías generales abstractas para contar con estimaciones estadísticas libres de sesgos.

8 Autores como Verger, Bonal y Zancajo (2016) señalan, por ejemplo, que las políticas educativas tienden a estar sustentadas en nociones racionalistas de las teorías de capital humano, según las cuales los estudiantes tienen un incentivo intrínseco a mejorar sus notas, ya que estas se correlacionan con su productividad y su ingreso futuro.

agregado). Es evidente que si nos acogemos al marco de referencia de la causalidad generativa propuesto por Pawson y Tilley (1997), esta promesa de los evaluadores experimentales carece de sustento metodológico. ¿Cómo es posible generar evidencias convincentes para informar procesos de decisión si, por definición, un ejercicio estadístico controlado no permite entender qué se hizo bien (o mal) y como (no) reproducirlo? Frente a dicho escenario, y sin recurrir a otras herramientas analíticas, la única conclusión que emerge de una evaluación experimental es que, para efectos de un debate público, una intervención X *podría*, *eventualmente*, estar correlacionada con un resultado Y (Krauss 2015). Por ende, insiste Cartwright (2013),

> la etiqueta de patrón de oro [aplicada] para […] los resultados de las RCT […] puede crear una falsa sensación de seguridad. Las RCT pueden ser el patrón de oro para afirmaciones de que la política funciona en la población estudiada en el entorno del estudio. Pero es solo una pieza de la evidencia necesaria para que suscriba la predicción de que [podría funcionar]. (101)

Para hablar del tipo de evidencia requerida para acercarse al objetivo del paradigma de la PBE y establecer algunos criterios que permitan evaluar su veracidad y relevancia, Cartwright y Hardie (2012) invitan al lector a pensar en la forma en que se resuelven crímenes en obras clásicas de la literatura policiaca. Con tal propósito, traen a colación el caso del inspector French del libro *Death on the Way* (de Freeman Wills Crofts), quien investiga la muerte de Ackerly, un personaje ficticio. Según narran, French consideraba contar, en un primer momento de la historia, con suficiente evidencia para incriminar a Carey, otro personaje inventado, como responsable de haber asesinado a Ackerly. Una primera premisa que sustentaba el razonamiento del inspector era que existían motivos para el crimen: Ackerly había denunciado un hecho fraudulento y existía soporte documental para sospechar que Carey fue su autor intelectual. Esta segunda hipótesis estaba sustentada en supuesta evidencia sobre la posición de poder de Carey para haber tenido la oportunidad de perpetrar el fraude y sobre la versión de que este se había suicidado una vez el escándalo se hizo público. Todo lo anterior indicaba que Carey pudo haber matado a Ackerly para silenciarlo, pero como ello no fue suficiente para ocultar el crimen, decidió también quitarse la vida. La narrativa del detective estaba además sustentada en detalles sobre cómo pudo haberse cometido el crimen, incluyendo lugares donde se encontraban los implicados y cálculos de tiempo de desplazamiento entre locaciones. Todo indicaba que French había identificado las piezas del rompecabezas. Su argumento era sofisticado y coherente. Pero al poco tiempo se revelaría un hecho que desbarataría

por completo su cadena de razonamiento: Carey no se había quitado la vida; fue asesinado. Dicho desenlace, sostienen Cartwright y Hardie (2012), abre paso a la siguiente reflexión:

> Entonces, toda la teoría cae. Ya no hay una razón convincente para pensar que Carey asesinó a Ackerly. ¿Pero qué hay de todas las pruebas que el francés tan cuidadosamente había reunido? Ya no era evidencia de que Carey fuera el asesino. El suicidio fue un soporte esencial para la suposición de que Carey era responsable del fraude. Sin el suicidio, la otra evidencia de esto era demasiado débil para justificar esa suposición. Pero sin eso, la premisa 1 falla; el motivo no está establecido. Y sin un motivo, la oportunidad y la existencia de un posible método para llevar a cabo el asesinato proporcionan una débil garantía, si es que hay alguna, de que Carey fue el asesino. Todavía es posible que fuera culpable. Pero French ya no tiene pruebas para sustentarlo. (20)

Danemark *et al.* (2002) hacen referencia a las ideas de Charles Sanders Peirce, precursor de la escuela pragmática norteamericana, para sugerir una analogía similar. El trabajo de un detective implica que las actividades de varias personas, las observaciones en campo y las afirmaciones hechas en entrevistas deben ser interpretadas y cobran significado de acuerdo con un marco de referencia o una hipótesis sobre cómo fue cometido un crimen.[9] Toda esta discusión permite entrever un hecho que puede sonar obvio, pero que a pesar de ello se escapa del sentido común básico en muchas discusiones sobre la PBE: la evidencia no representa una propiedad intrínseca de un dato, un indicador, un estimador estadístico o una anécdota que recupera el investigador durante el trabajo de campo. Por tanto, tiene poco sentido hablar de *bancos de evidencias* para referirnos a un cuerpo neutral de información que reposa en un archivo, una página de internet o en las bases de datos de agencias nacionales e internacionales (Munro *et al.* 2016). X solo puede servir de envidencia de Y cuando alguien toma la decisión de que X es relevante para su proceso de deliberación sobre los mecanismos detrás de la existencia de Y. Puesto en términos algo más formales, "la relevancia evidencial es entonces una relación de tres elementos. Implica una declaración de [que X es] evidencia [de Y], una hipótesis (o conclusión) y un argumento" (Cartwright y Hardie 2012, 18).

9 Puesto en términos formales, esto corresponde a un ejercicio de abducción. Para una explicación más detallada de este concepto, consultar Parra (2019).

Nada de lo anterior supone dejar a un lado criterios de rigor metodológico (por ejemplo, la solidez estadística de un modelo econométrico o la triangulación de información cuantitativa y cualitativa para corroborar afirmaciones de diferentes partes entrevistadas) de una evaluación. Pero, nuevamente, dichos criterios técnicos no suplantan la consistencia de un argumento. Usemos otro ejemplo de Cartwright (2017) para examinar posibles principios orientados a valorar el buen uso de la evidencia al momento de estudiar una intervención social. La filósofa nos invita a imaginar a un individuo *i* que, sin proponérselo, ingirió una sustancia venenosa (*v*). Por fortuna, *i* cayó en la cuenta del error con cierta anticipación y decidió consumir un emético (*m*) que, como es previsible (por sus propiedades medicinales), le produjo ganas de vomitar. Tras un examen con un especialista, *i* se enteró de que su cuerpo no alcanzó a sufrir síntomas serios de evenenamiento. Es tendador concluir, por ende, que tenemos evidencia para inferir que el remedio tuvo un efecto causal en contener la intoxicación. El cuadro 1 presenta una serie de elementos que nos sirven para llevar a cabo dicha valoración. Luego de examinarlo con detalle podremos hablar de criterios más generales que contribuyan a orientar a evaluadores profesionales en diferentes escenarios.

Cuadro 1. Validación de la evidencia sobre efectos de un medicamento

a.	**Eliminación de alternativas:** la tasa de supervivencia con este veneno es muy baja. Por lo tanto, no es probable que mi supervivencia fuera espontánea. Y no tengo ninguna característica especial en mi organismo que explique mi supervivencia tras haber consumido el veneno: no tengo una masa corporal excepcional, no me había aclimatado lentamente a este veneno con dosis menores anteriores, no tomé un antídoto, etc.
b.	**Presencia de factores de soporte necesarios, sin los cuales no se puede esperar que el veneno tenga efecto:** el emético se ingirió antes de que se absorbiera demasiado veneno en el estómago.
c.	**Presencia del paso intermedio necesario:** vomité.
d.	**Presencia de síntomas de las causas putativas que actúan para producir el efecto:** había mucho veneno en el vómito, que es un claro efecto secundario de que los eméticos son los responsables de mi supervivencia.
e.	**Características del efecto:** se midió la cantidad de veneno en el vómito y se comparó con la cantidad que había consumido. Sufrí solo los efectos de la cantidad restante de veneno; y el momento del efecto y el tamaño fueron correctos.

Fuente: adaptado de Cartwright (2017).

No vamos a detallar el contenido del cuadro para evitar ser reiterativos. Nos centramos en lo que este devela frente a la importancia de contar con el apoyo de información que sirva tanto de *evidencia directa* como de *evidencia*

indirecta para dar soporte a una hipótesis causal: el medicamento (o emético) *m* tuvo incidencia en reducir el efecto del veneno *v* en el cuerpo del individuo *i*. La evidencia directa representa todo tipo de información que haga referencia a la relación causal primaria que quiere ser explicada. De acuerdo con Cartwright (2017), esta puede hacer alusión a las características del efecto y cómo se espera —según, por ejemplo, literatura especializada— que este se dé en el cuerpo humano (es decir, *a* en el cuadro 1). De igual forma, resulta oportuno preguntarse sobre los síntomas de la posible causalidad, incluyendo la valoración de efectos secundarios que podrían esperarse del proceso en que se genera el resultado final (*d* en el cuadro 1), sobre la presencia de factores de soporte (concepto en el que ahondamos más adelante en el texto) que indiquen la existencia de condiciones necesarias (internas al proceso) para que *m* tuviera un impacto en *v* (*b* en el cuadro 1) o la identificación de pasos intermedios en la cadena causal de eventos (*e* en el cuadro 1). La evidencia indirecta, por su parte, nos ayuda a reflexionar sobre la posibilidad de que nuestra hipótesis causal sea cierta. Cartwright (2017) habla, por tanto, de información que nos sirve para eliminar alternativas a la explicación del surgimiento del fenómeno estudiado (el paso de *m* a reducir el efecto de *v*). La frase emblemática de Sherlock Holmes que indica que "cuando eliminas toda solución lógica a un problema, lo ilógico, aunque imposible, es invariablemente lo cierto" resulta pertinente para ilustrarlo (ver segunda frase de *a* en el cuadro 1).

Insistamos con el mensaje general: "[e]s importante tener claro cuáles son los argumentos para sustentar [una] conclusión y [valorar] qué tan bien respaldadas están sus premisas para evaluar qué tan seguros [podemos] estar de que la conclusión es correcta" (Cartwright y Hardie 2012, 19). Criterios como los esbozados en el cuadro 1 ayudan a aterrizar dicha premisa de la filosofía a un plano metodológico. Surge en este punto una pregunta legítima frente a la existencia de criterios cerrados que sirvan al investigador para valorar si cuenta con la cantidad suficiente de evidencia directa e indirecta que respalde su conclusión causal. Este interrogante nos permite enlazar esta discusión con la distinción entre tipos de causalidad planteada en la primera sección para decir lo siguiente: todo depende de las preferencias metodológicas del evaluador. Un camino sucesionista reflejado, nuevamente, en el uso primario de RCT (u otros métodos experimentales) hallaría criterios de rigurosidad en la aplicación de pruebas estadísticas que verifiquen la consistencia interna de un modelo econométrico contrafactual. Sin embargo, como se ha sugerido previamente, este camino es útil tan solo para verificar que *m* está definitivamente asociado al efecto de *v* en *i*, y no para entender el cómo ni el porqué de esa asociación. De otro lado, una alternativa fundamentada en la noción de causalidad generativa privilegiaría más

la exploración de las causas eficaces, esquema bajo el cual cobra importancia el estudio de las motivaciones humanas para actuar de una manera y no de otra. En esta segunda lógica resulta difícil para el investigador establecer si cuenta con toda la información necesaria para sustentar la totalidad de su razonamiento y, por tanto, solo le queda recurrir a un principio general: más evidencia es mejor (Cartwright 2017).

Esta última conclusión podría resultar desilusionante (o abstracta) para aquellos interesados en aprender sobre cómo mejorar el funcionamiento de diferentes políticas o programas sociales. Existen, no obstante, rutas metodológicas alternativas —o, si se quiere, complementarias— a los esquemas experimentales, y que parecen prometedoras frente a la definición de criterios más específicos para la valoración de la evidencia sobre procesos causales. La *evaluación realista* es un buen ejemplo, y su creciente influencia en el trabajo de evaluadores en diferentes sectores y temáticas así lo sugiere (Emmel *et al*. 2018; Jagosh, Tilley y Stern 2016). Por motivos de espacio y enfoque del texto, no se profundizará en esta tradición y su apuesta explícita frente a la aplicación de lógicas causales generativas en diseños de investigación (Brouselle y Burejeya 2018; Pawson, Wong y Owen 2011). Textos introductorios como el del Parra (2017) ofrecen una síntesis de sus principios rectores.

3. La validez externa y el reto de extrapolar poderes causales eficaces

Según Deaton y Cartwright (2018),

> [el] concepto binario de validez externa es a menudo inservible porque les pide a los resultados de una RCT satisfacer una condición que no es necesaria ni suficiente […] Nos dirige hacia una simple extrapolación —ya sea [para sugerir] que el mismo resultado es válido en otra parte— o a una simple generalización —[un hallazgo] se mantiene universalmente o al menos ampliamente—, y nos aleja de aplicaciones más complejas pero más útiles de los resultados […] Sin ir más allá en la comprensión y el análisis, incluso una reproducción exitosa [de un instrumento de política] nos dice poco […] para apoyar la conclusión de que la próxima vez esta va a funcionar de la misma manera. (10)

Hasta el momento hemos centrado la reflexión en interrogantes sobre la identificación de evidencia que permita concluir que una política o programa evaluado pudieron haber funcionado en un contexto específico (que fue evaluado) y

por qué. La discusión ha reconocido el papel de las evaluaciones experimentales como instrumento para confirmar que, al menos en el plano empírico (u observacional), dos variables (una de tratamiento y otra de resultado) se relacionan entre sí (y también qué tanto se relacionan). Al tiempo, se reitera la función limitada de estas técnicas en la generación de evidencia tanto directa como indirecta sobre la senda causal entre la implementación de una medida, su recepción por parte de un grupo de beneficiarios y el curso de acción eficaz de estos con respecto a los incentivos que les son ofrecidos (un subsidio, una capacitación, un mayor acceso a un sistema de crédito, etc.). La discusión en torno a la pertinencia del concepto de validez externa se desprende, como un posible corolario, de este reconocimiento de las limitaciones de la estadística para identificar relaciones causales útiles para pensar acerca del diseño de una intervención social. Para Danemark *et al.* (2002), "el estudio de regularidades empíricas, o de una covariación entre variables estandarizadas, no puede sustentar opiniones sobre nada más que regularidades empíricas y correlaciones estadísticas; no pueden responder preguntas acerca de las causas" (53). Si aceptamos este argumento, ¿tiene sentido hablar de validez externa de una *explicación* estadística? La respuesta simple es no; no tiene sentido extrapolar algo que no existe (Krauss 2015).

Se pueden contemplar, sin embargo, rutas alternativas. Para discutirlas resulta relevante traer a colación otros elementos para matizar (o aclarar) la conclusión enunciada en el párrafo anterior. Esto es importante en tanto, al asumirla como una refutación absoluta de la factibilidad de extrapolar conocimientos sobre mecanismos sociales, se dilapidaría la posibilidad de la PBE. De vuelta a debates ya enunciados en este artículo, para que un método como las RCT permita generar conocimiento aplicable a múltiples escenarios, estas deberían tener la capacidad de identificar no solamente condiciones necesarias (como entornos institucionales, materiales, culturales), sino también poderes causales eficaces (como acciones e intenciones humanas). Esta es una forma diferente de referirse al concepto de causalidad generativa previamente discutido y que, hemos insistido, difiere de los principios que orientan prácticas dominantes de evaluación De hecho, según resaltan Deaton y Cartwright (2018), la ruta más común de los evaluadores experimentalistas es asumir el modelo de actor racional propuesto en la teoría económica convencional y, tras ello, equiparar lo necesario (por ejemplo, *buenos* incentivos) con la conducta racional (por ejemplo, una que responda a dichos incentivos). En palabras de estos autores:

> [La tradición] de pensar sobre los cambios de política como algo equivalente a los cambios del precio y el ingreso tiene una larga historia en economía; gran parte de la teoría de la elección racional puede ser así

> interpretada […] Cuando esta conversión es creíble, y cuando un ensayo sobre un tema aparentemente no relacionado puede ser modelado como equivalente a un cambio en los precios y los ingresos, y cuando podemos suponer que las personas en diferentes entornos responden de manera similar a cambios en los precios y los ingresos, tenemos un marco listo para incorporar el resultado de las pruebas en [el acervo de] conocimientos previos, así como para […] usarlos en otra parte. (15)

Bajo dicha ruta metodológica, habría pocas razones para problematizar la noción de validez externa en tanto el mismo diseño (racionalista) de una intervención social y la aplicación de elementos básicos de la teoría neoclásica del consumidor garantizarían el poder contar con buenas predicciones sobre lo que sucedería si se utiliza lo que aprendemos en un contexto X en un escenario Y. No obstante, los economistas del comportamiento se han encargado de desmitificar el supuesto del hombre-económico que se requiere para legitimar ese tipo de argumentación. De acuerdo con un influyente estudio de experimentos sociales en quince sociedades realizado hace casi dos décadas,

> el modelo canónico del actor egoísta maximizador de pagos materiales se viola sistemáticamente. En todas las sociedades estudiadas, las ofertas de [un juego diseñado para evaluar el altruismo] son estrictamente positivas y a menudo sustancialmente superiores a la oferta de maximización de ingresos esperada […] El resultado implica que los juicios de la economía del bienestar que asumen preferencias exógenas son cuestionables, al igual que las predicciones de los efectos de cambios en políticas económicas y las instituciones que no tienen en cuenta cambios de tipo comportamental. (Henrich *et al*. 2001, 77)

El concepto de preferencias endógenas discutido por autores como Bowles (2008) cobra gran relevancia en este contexto en tanto, como su nombre lo anuncia, entra en conflicto con la noción de variación *exógena* que defienden, como criterio de rigurosidad técnica, los exponentes de esquemas de evaluación experimental.[10] A diferencia de una noción de exogeneidad, la cual evoca la pretensión de la economía positiva de analizar a los seres humanos *tal y como son*,[11] la noción de endogeneidad coincide con una visión donde los mercados y otras

10 Según Pawson y Tilley (1997), los métodos de evaluación sucesionistas basan su razonamiento en la búsqueda de una variación exógena, o externa, que, según dicho paradigma, garantice una estimación libre de sesgos estadísticos (o de endogeneidades).

11 En términos ontológicos esto equivale a argumentar que las personas tienen preferencias fijas en el tiempo.

instituciones políticas, sociales y económicas "hacen más que distribuir bienes y servicios: también influyen en la evolución de valores, los gustos y las personalidades" (Bowles 2008, 75). Dicho de otra manera, si una política pública tiene la capacidad de incidir en las preferencias de un individuo o grupo social, más allá de simplemente direccionarlos por la senda de lo *racional*, resulta incongruente asumir *a priori* que un mismo instrumento, como un subsidio o un beneficio tributario, tendrá siempre el mismo resultado sobre su población objetivo.[12]

Este último argumento es quizás más evidente cuando se piensa en extender la implementación de una política pública sobre una misma población. Pero ¿y qué pasa cuando se piensa en extrapolar los resultados de un estudio de un contexto a otro? La intuición dicta que el problema se torna más complejo. Y en efecto lo es. Desde una teoría generativa de la causalidad, según la cual existen circunstancias o condiciones específicas que llevan a que dos eventos sociales guarden una relación causal entre sí (por ejemplo, las habilidades de los docentes y los resultados de aprendizaje de un estudiante),[13] las características (o propiedades) del contexto donde se despliega una política o programa serán determinantes en el cumplimiento, o no, de sus objetivos. Por ejemplo, si bien en algunas culturas el enviar a padres de familia a clases sobre la crianza puede incidir en el mejoramiento del comportamiento de sus hijos, en otras, quizas más machistas, los hombres pueden sentirse humillados públicamente y reaccionar violentamente contra miembros de su hogar (Munro *et al.* 2016). En el caso del manejo macroeconómico de un país, puede que un aumento en la oferta monetaria no incida en un incremento de la inflación, según predice el monetarismo, debido a arreglos institucionales específicos (como la confianza en la regulación bancaria) bajo los cuales interactuan sus ciudadanos (Reiss 2018). Razonamientos similares pueden aplicarse al campo de la medicina, en el cual es bien sabido que factores sociales y culturales inciden en el uso, y el tipo de uso, de medicamentos por parte de la población (Deaton y Cartwright 2018).

Nuevamente, ¿quiere esto decir que no se puede, bajo ninguna circunstancia, extraer lecciones del funcionamiento de políticas públicas o proyectos sociales en un contexto para aplicarlas en otro? Aceptar dicho postulado sería equivalente a negar la posibilidad de evaluar intervenciones sociales. Para Cartwright y Hardie (2017) no se trata, sin embargo, de un simple dilema binario; el debate, en línea

12 He hecho una explicación sucinta de este razonamiento en Parra (2013).

13 En su revisión de experiencias internacionales sobre los determinantes del aprendizaje en educación básica y secundaria, Masino y Niño-Zarazúa (2016) concluyen que en contextos donde el servicio educativo se encuentra condicionado por situaciones sociales y económicas específicas, el uso de capital humano (por ejemplo, excelencia docente), por sí mismo, puede resultar en una pérdida de dinero.

con lo discutido en el capítulo anterior, gira en torno a cómo generar la mejor evidencia posible, dada la complejidad intrínseca en todo intento de predecir el comportamiento de las personas. En su libro *Política basada en evidencia: una guía práctica para hacerlo mejor*, estos autores nos invitan a reexaminar el problema a partir de un razonamiento lógico que surge del reconocimiento de las limitaciones de las aproximaciones dominantes (por ejemplo, las RCT) para el estudio de la sociedad. Este puede darse a partir del planteamiento de dos interrogantes concretos que debería hacerse todo aquel interesado en aprender de los hallazgos sobre el funcionamiento de una intervención social en un contexto para aplicarlos en otro:

i. ¿Tiene la intervención la capacidad *real* de ayudarnos a resolver un problema específico?
ii. ¿Existen factores de soporte que garanticen que algo que funcionó en el contexto Y pueda funcionar en X?

La primera pregunta tiene que ver con la reflexión en torno al comportamiento esperado de posibles beneficiarios de la expansión de una intervención. Nuevamente, las intervenciones sociales no funcionan porque sí; existen razones específicas (como la acción de mecanimos sociales concretos) detrás de sus resultados. El problema con esquemas de evaluación como los de las RCT, es imporante insistir, es que no permiten a los evaluadores contar con buenas teorías sobre esas razones comportamentales específicas y sobre cómo operan en diferentes contextos. Un posible corolario de ello es que el primer paso lógico para pensar en extrapolar enseñanzas de un contexto a otro consiste en utilizar estrategias de evaluación fundamentadas en visiones generativas (en lugar de sucesionistas) de la causalidad. Nuevamente, las metodologías de las evaluaciones basadas en teoría (Brouselle y Burejeya 2018) representan una alternativa relevante para emprender esta tarea (Cartwright 2013; Deaton 2010; Pawson 2013), en tanto su objeto de análisis es, precisamente, estudiar la forma en que diferentes mecanismos, operando en diferentes sistemas sociales, pueden explicar el surgimiento de diferentes eventos sociales (como el bajo desempeño escolar, la reducción del crimen, el uso de medicamentos).

El segundo interrogante hace referencia al contexto específico en el cual se busca extrapolar enseñanzas de otras experiencias. Una vez tengamos una idea de que una intervención cuenta con la capacidad (potencial) para fomentar un cambio deseado, debemos reflexionar sobre si las características del nuevo contexto pueden ser conducentes a la activación de los mecanismos causales esperados. Esas condiciones son lo que Cartwright y Hardie (2012) denominan *factores de soporte*. Se contemplan acá desde elementos materiales concretos, como la existencia de alternativas de transporte asequibles para que padres de familia lleguen a la locación física donde van a recibir orientaciones sobre la crianza de sus hijos, hasta elementos culturales (convicciones y convenciones

sociales, relaciones entre hombres y mujeres, etc.) que puedan afectar las decisiones de un individuo o de un grupo de personas. Desde luego, pensar sobre posibles factores de soporte implica márgenes de incertidumbre que, no obstante, pueden afrontarse (hasta cierto punto) a partir de la valoración de hallazgos que surjan del análisis (en la primera pregunta) sobre mecanismos y contextos detrás del potencial de una política o proyecto para *hacer algo*.

Una breve ilustración sobre la aplicación de dichos criterios puede servir para aclarar implicaciones de estos últimos mensajes. Cartwright y Hardie (2012) toman como ejemplo el caso del programa de reducción del tamaño de clase aplicado en el estado de California, en Estados Unidos, a mediados de los años noventa. Se trata del intento de poner en práctica una política pública exitosa en otro estado, Tennessee, la cual, según resultados de una RCT, logró mejorar el desempeño escolar de estudiantes en salones más pequeños y con una menor relación docente-alumno. Mencionan los autores que dicha política no solo había pasado la prueba ácida de una evaluación de impacto rigurosa, sino que también dialogaba con cierto sentido común que confirmaba la ideoneidad de la política. Pese a ello, una nueva RCT en 2002, igual de rigurosa que la del primer estudio, arrojó resultados no significativos que confirmaban la inexistencia de cualquier vínculo causal entre la política de reducción de tamaño de la clase y el mejoramiento de los resultados en exámenes de estudiantes californianos. ¿Qué explica dicho resultado? Según Cartwright y Hardie (2012), el experimento en el segundo contexto no arrojó los resultados esperados por la ausencia de dos factores de soporte: la disponibilidad de espacio y la falta de docentes preparados:

> En Tennessee, el proyecto involucró solo a las escuelas que tenían espacio disponible. En California, a menudo no había suficiente espacio libre. A veces se encontró espacio, pero no era tan bueno como los salones de clase existentes. Y [el espacio] fue retirado de otras actividades que podrían considerarse importantes para el rendimiento estudiantil: necesidades especiales, música y artes, deportes y programas de cuidado infantil. En segundo lugar, Tennessee no tenía escasez de maestros calificados para el personal de las clases de tamaño reducido. Pero California tuvo que contratar a 12.000 maestros adicionales. Y muchos de estos no estaban calificados. (65)

A la luz de los dos interrogantes discutidos arriba, el fracaso del programa de reducción de clases en California se puede entender, en parte, por la inexistencia de una valoración profunda de las razones causales reales detrás del éxito de la experiencia del experimento en Tennessee. Si bien los autores del ejemplo no entran en detalles sobre esta primera discusión, sí mencionan elementos como la

búsqueda de espacios (físicos y de tiempo) para abrir más aulas de clase y cómo ello implicó que estudiantes californianos perdieran horas dedicadas al trabajo en materias como música o artes. Quizas si la RCT inicial hubiese hecho alguna previsión metodológica para estudiar (y no solo asumir) mecanismos comportamentales, sus ejecutores hubieran identificado que tocar un instrumento musical o la expresión artística pueden afectar la autoestima y la motivación de niñas y niños y, así, impactar su desarrollo cognitivo, tal y como lo demuestra la literatura (Hallam 2010). Si ese hubiese sido el caso (acá nos movemos en un terreno hipotético), los investigadores o hacedores de política interesados en extrapolar enseñanzas de un contexto a otro habrían previsto que la reducción del tamaño del aula tiene *la capacidad* de mejorar el desempeño académico si, y solo si, ello no va en detrimento del deseo de los estudiantes de aprender.

De otro lado, en el ejemplo es claro que el programa en el segundo caso falló porque el sistema educativo de California no contaba con factores de soporte tanto físicos (docentes capacitados, espacios físicos apropiados para la enseñanza) como culturales (por ejemplo, el deseo de estudiar que se puede desatar del aprendizaje no cognitivo de los estudiantes) para que se reprodujera la teoría de cambio que fue exitosa en Tennessee. Desde luego, y pese a que la aplicación de herramientas metodológicas más informadas por una lógica generativa de la causalidad hubiera permitido contar con un mejor conocimiento (frente a la opción experimentalista) de esos factores de soporte, estos no podrían haber sido identificados *a priori* con un grado de certeza estadística. Por tal razón, exaltan Cartwright y Hardie (2017), "[i]ncluso cuando hemos hecho lo mejor [posible], la lección básica permanece. La vida es complicada y los resultados son difíciles de predecir, incluso si no adoptamos un modelo de causalidad dominó simple y tratamos de delinear un modelo más realista" (272). En últimas, cierto grado de prueba y error es inevitable cuando se entra en el campo complejo de las intervenciones sociales (Monaghan, Pawson y Wicker 2012). Dictamina Reiss (2018), por ende, que "una comunidad epistémica que pone la certeza antes de la relevancia no obtendrá ninguna" (13). Por el contrario, sugiere, "una comunidad epistémica que pone la relevancia antes de la certeza obtendrá un alto grado de ambos" (13). Esta es una forma sofisticada de poner en el contexto de debates sobre la filosofía del conocimiento en las ciencias sociales las lecciones metodógicas que se derivan del fallido experimento californiano.

Conclusión

En los debates sobre la denominada PBE se tiende a asumir, como punto de partida, la jerarquía absoluta de los métodos experimentales de evaluación de impacto como garantes de rigor y objetividad. Las RCT representan, por tanto, una

especie de patrón de oro para informar a hacedores de politica y administradores de proyectos sociales frente a cómo y dónde invertir recursos para generar transformaciones deseables (por ejemplo, mejorar el desempeño escolar, reducir tasas de criminalidad, incrementar índices de salud). El problema con dicho razonamiento, según discutimos en el texto, es que las evaluaciones experimentales no cuentan con el aparato conceptual o metodológico necesario para ayudarnos a entender por qué y cómo una intervención social, en un contexto específico, pudo llevar, o no, al cumplimiento de sus objetivos. Si bien la econometría aplicada puede ser muy relevante para constatar que dos fenómenos sociales (como un subsidio al desempleo y la variación en tasas de ocupación) están efectivamente correlacionados, no dice nada sobre la dinámica social concreta que permitió que existiese esa correlación. Sin ese tipo de conocimiento (o evidencia), no solo se reduce la probabilidad de reproducir un mismo resultado en un mismo contexto (por ejemplo, expandir una política en una ciudad por unos años adicionales), sino que pierde todo sentido hablar de validez externa (o la posibilidad de replicar resultados en otros contextos), al menos en la forma casi automática que pregonan los evaluadores experimentalistas.

La segunda parte del documento se adentra en explorar posibles alternativas metodológicas frente a las limitaciones de los enfoques sucesionistas de la evaluación al momento de ayudarnos a predecir qué podría pasar si intervenimos en un lugar o una población específica. La noción de predicción a la que se hace referencia no puede ser estadística, por lo discutido arriba, sino más de corte cualitativo. Atendiendo a un razonamiento lógico sobre cómo enmendar los problemas de las prácticas dominantes de la evaluación, el artículo propone a los interesados en evaluar intervenciones sociales hacerse dos preguntas: i) ¿tiene esta intervención la capacidad real de generar el resultado X?; y ii) si queremos extrapolar los hallazgos de un contexto a otro, ¿existen los factores de soporte (en el nuevo contexto) para esperar que dicha (posible) capacidad de la política o proyecto efectivamente se traduzca en una transformación social deseada? Para poder responderlas, se invita a explorar propuestas desde la tradición de las evaluaciones basadas en teoría, en tanto estas se han propuesto como objetivo estudiar, precisamente, cómo y por qué funcionan las políticas o proyectos sociales en diferentes contextos. Algunos de los principios básicos sobre cómo valorar evidencia (de algo) discutidos en el artículo pueden también ayudar a mejorar nuestro entendimiento sobre dinámicas sociales complejas.

Como comentario final, quisiera insistir en lo incierto que es trabajar o analizar ambientes o ámbitos humanos y, por tanto, en la dificultad intrínseca que implica la idea de predecir qué podría pasar. Esto no con la intención de sembrar un ambiente de pesimismo sobre la capacidad trasnformadora de diferentes políticas

y proyectos sociales. Todo lo contrario; tengo la convicción de que si asimilamos que es simplemente imposible predecir (al menos con exactitud probabilística) qué va a pasar, nuestros esfuerzos por generar mejor evidencia van a aumentar. Esta es una invitación a la deliberación más profunda sobre las diferentes dimensiones de las intervenciones sociales, la cual debe además servir de plataforma para que más profesionales de las ciencias sociales (y no solo, como es común, estadísticos, ingenieros y economistas) puedan aportar, y contar con la suficiente legitimidad, para evaluarlas. Habría que agregar que dicha incertidumbre debe ser también motivo para mejorar las prácticas de monitoreo (en el corto y mediano plazo) de las intervenciones sociales, de modo que no solo los números o los indicadores duros tengan un papel protagónico. Si los seres humanos son, en últimas, la causa eficaz de la sociedad, monitorear implica también observar, conversar y estudiar sus percepciones, y como cambian en el tiempo.

Referencias

1. Bernal, Raquel y Ximena Peña. 2011. *Guía práctica para la evaluación de impacto*. Bogotá: Ediciones Uniandes.
2. Blamey, Avril y Mhairi Mackenzie. 2007. "Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges?". *Evaluation* 13 (4): 439-455.
3. Bowles, Samuel. 2008. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions". *Journal of Economic Literature* 36 (1): 75-111.
4. Brouselle, Astrid y Jean-Marie Burejeya. 2018. "Theory-based Evaluations: Framing the Existence of the New Theory Evaluation and the Rise of the 5th Generation". *Evaluation* 24 (2): 153-168.
5. Cardozo, Myriam. 2013. "De la evaluación a la reformulación de políticas públicas". *Política y Cultura* 40: 123-149.
6. Cartwright, Nancy. 2012. "Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philoshophy Helps". *Philosophy of Science* 79 (5): 973-989.
7. Cartwright, Nancy. 2013. "Knowing What We Are Talking About: Why Evidence Doesn't Always Travel". *Evidence & Policy* 9 (1): 97-112.
8. Cartwright, Nancy. 2017. "Single Case Causes: What is Evidence and Why". En *Philosophy of Science in Practice: Nancy Cartwright and the Nature of Scientific Reasoning*, editado por Hsiang-Ke Chao y Julian Reiss, 11-24. Cham: Springer.
9. Cartwright, Nancy y Jeremie Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.
10. Cartwright, Nancy y Jeremie Hardie. 2017. "Predicting What Will Happen When You Intervene". *Clinical Social Work Journal* 45 (3): 270-279.
11. Danermark, Berth, Mats Ekström, Liselotte Jakobsen y Jan Ch Karlsson. 2002. *Explaining Society: Critical Realism in the Social Sciences*. Nueva York: Psychology Press.
12. Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development". *Journal of Economic Literature* 48: 424-455.

13. Deaton, Angus y Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials". *Social Science & Medicine* 210: 2-21.
14. Ellis, George. 2005. "Physics, Complexity and Causality". *Nature* 435 (743).
15. Gertler, Paul J., Sebastián Martínez, Patrick Premand, Laura B. Rawlings y Christel M. J. Vermeersch. 2011. *Evaluación de impacto en la práctica.* Washington D. C.: Banco Mundial.
16. Goodman, Lisa, Deborah Epstein y Cris Sullivan. 2018. "Beyond the RCT: Integrating Rigor and Relevance to Evaluate the Outcomes of Domestic Violence Programs". *American Journal of Evaluation* 39 (1): 58-70.
17. Greenhalgh, Trish y Craig Russell. 2009. "Evidence-Based Policymaking: A Critique". *Perspectives in Biology and Medicine* 52 (2): 304-318.
18. Hallam, Susan. 2010. "The Power of Music: Its Impact on the Intellectual, Social and Personal Development of Children and Young People". *International Journal of Music Education* 28 (3): 269-289.
19. Harman, Graham. 2018. *Object-Oriented Ontology.* Londres: Pelican Books.
20. Hausmann, Ricardo. 2016. "El problema con las políticas basadas en evidencia". *La Nación*, 3 de marzo. https://www.nacion.com/opinion/foros/el-problema-con-las-politicas-basadas-en-evidencia/ISW3U6CZYNHEVND43YO2XZWFBY/story/
21. Head, Brian. 2016. "Toward More 'Evidence-Informed' Policy Making?". *Public Administration Review* 76 (3): 472-484.
22. Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis y Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies". *American Economic Review* 91 (2): 73-78.
23. Joyce, Kathryn y Nancy Cartwright. 2020. "Bridging the Gap Between Research and Practice: Predicting What Will Work Locally". *American Educational Research Journal* 57 (3): 1045-1082.
24. Krause, Philipp y Gonzalo Hernández. 2020. "Commentary: From Experimental Findings to Evidence-Based Policy". *World Development* 127. https://doi.org/10.1016/j.worlddev.2019.104812
25. Krauss, Alexander. 2015. "The Scientific Limits of Understanding the (Potential) Relationship Between Complex Social Phenomena: The Case of Democracy and Inequality". *Journal of Economic Methodology* 23 (1): 97-109.
26. Krauss, Alexander. 2018. "Why All Randomised Controlled Trials Produce Biased Results". *Annals of Medicine* 50 (4): 312-322.
27. Lawson, Tony. 2009. "Applied Economics, Contrast Explanation and Asymmetric Information". *Cambridge Journal of Economics* 33 (3): 405-419.
28. Manzano, Ana. 2010. "El análisis del contexto local en un programa multidisciplinario (sanidad y servicios sociales) usando el enfoque de la evaluación realista". *E-valuacion* 3 (10): 24-27.
29. Masino, Serena y Miguel Niño-Zarazúa. 2016. "What Works to Improve the Quality of Student Learning in Developing Countries?". *International Journal of Educational Development* 48: 53-65.
30. Monaghan, Mark, Ray Pawson y Kate Wicker. 2012. "The Precautionary Principle and Evidence-Based Policy Making". *Evidence & Policy* 8 (2): 171-191.
31. Munro, Eileen, Nancy Cartwright, Jeremy Hardie y Eleonora Montuschi. 2016. *Improving Child Safety: Deliberation, Judgement and Empirical Research.* Durham: Centre for Humanities Engaging Science and Society (Chess). Philosophy Department, Durham University.

32. Parra, Juan David. 2013. "Preferencias endógenas, prosocialidad y políticas públicas". *Divergencia* 15: 64-71.
33. Parra, Juan David. 2016. "Realismo crítico: una alternativa en el análisis social". *Sociedad y Economía* 31: 215-238.
34. Parra, Juan David. 2017. "¿Qué funciona, para quién, en qué aspectos, hasta qué punto, en qué contexto y cómo? Una introducción a la evaluación realista y sus métodos". *Economía & Región* 11 (2): 11-44.
35. Parra, Juan David. 2018. "Critical Realism and School Effectiveness Research in Colombia: The Difference It Should Make". *British Journal of Sociology of Education* 39 (1): 107-125.
36. Parra, Juan David. 2019. "El arte del muestreo cualitativo y su importancia para la evaluación y la investigación de políticas públicas: una aproximación realista". *Opera* 25: 119-136.
37. Pawson, Ray. 2013. *The Science of Evaluation: A Realist Manifesto*. Londres: Sage.
38. Pawson, Ray y Nick Tilley. 1997. *Realistic Evaluation*. Londres; Nueva Delhi: Thousand Oaks.
39. Pawson, Ray, Geoff Wong y Lesley Owen. 2011. "Known Knowns, Known Unknowns, Unknown Unknowns: The Predicament of Evidence-Based Policy". *American Journal of Evaluation* 32 (4): 518-546.
40. Porter, Sam, Tracey McConnell y Joanne Reidcor. 2017. "The Possibility of Critical Realist Randomised Controlled Trials". *Trials* 18: 133.
41. Reiss, Julian. 2018. "Against External Validity". *Synthese* 196 (8): 3103-3121.
42. Saltelli, Andrea y Mario Giampietro. 2017. "What Is Wrong with Evidence Based Policy, and How Can It Be Improved?". *Futures* 91: 62-71.
43. Van Belle, Sara, Geoff Wong, Gill Westhorp, Mark Pearson, Nick Emmel, Ana Manzano y Bruno Marchal. 2016. "Can 'Realist' Randomised Controlled Trials Be Genuinely Realist?". *Trials* 17 (1): 2-6.
44. Verger, Antoni, Xavier Bonal y Adrián Zancajo. 2016. "What Are the Role and Impact of Public-Private Partnerships in Education? A Realist Evaluation of the Chilean Education Quasi-Market". *Comparative Education Review* 60 (2): 223-248.

Juan David Parra es profesor asistente del Instituto de Estudios en Educación (IESE) de la Universidad del Norte. Es profesional en Gobierno y Relaciones Internacionales de la Universidad Externado; cursó el Doctorado en Estudios sobre el Desarrollo en ISS-Erasmus University Rotterdam y la Maestría en Economía de la Universidad de los Andes. En su acercamiento a la práctica en evaluación de programas y políticas públicas, destaca su paso por la firma Econometría Consultores, con la cual participó en proyectos nacionales e internacionales de diversos sectores económicos y sociales. Su interés académico actual se centra en la aplicación de metodologías *realistas* en campos como la evaluación y la economía política de la educación. Ha publicado trabajos empíricos y teóricos, principalmente en revistas especializadas como *British Journal of Sociology of Education*, *Third World Quarterly*, *International Journal of Qualitative Studies in Education*, *Revista Brasileira de Educação* y *Opera*, entre otras. jparrad@uninorte.edu.co