INTRODUCCIÓN A LA EVALUACIÓN REALISTA Y SUS MÉTODOS: ¿QUÉ FUNCIONA, PARA QUIÉN, EN QUÉ ASPECTOS, HASTA QUÉ PUNTO, EN QUÉ CONTEXTO Y CÓMO?

JUAN DAVID PARRA*

RESUMEN

Este artículo introduce la Evaluación Realista (ER) para el análisis de políticas públicas. La ER encuentra justificación en las limitaciones de prácticas dominantes, como las técnicas de evaluación de impacto, para ofrecer explicaciones sobre el por qué una intervención o programa tuvo (o no) resultado. La noción de cambio o transformación de la ER se centra en el estudio contextual de la agencia humana, o la forma en que individuos y colectividades actúan frente a intervenciones públicas o privadas. El texto se adentra en propuestas metodológicas para el despliegue de una ER en la práctica, tarea que implicará un trabajo interdisciplinario para refinar diferentes teorías sobre qué funciona, para quién, en qué aspectos, hasta qué punto, en qué contexto y cómo. La introducción del texto incluye una breve discusión sobre el potencial de una propuesta metodológica de este tipo para fortalecer el diseño y la implementación de iniciativas de política

* Juan David Parra es investigador del International Institute of Social Studies, Erasmus University Rotterdam. Correo electrónico: jparrah@gmail.com. El autor ofrece un agradecimiento especial a Ana Manzano, de la Universidad de Leeds (UK), por sus comentarios a una versión preliminar de este artículo. Las ideas expuestas en el texto fueron también alimentadas tras valiosos diálogos con otros evaluadores realistas experimentados como Geoff Wong y Ray Pawson. El autor también hace un reconocimiento especial a los integrantes del curso corto *Realist Reviews and Realist Evaluations* que se llevó a cabo en la Universidad de Oxford en Octubre de 2016. Los posibles errores en la interpretación de los conceptos de la ER son responsabilidad única del autor. Recibido: julio 30 de 2017; aceptado: noviembre 29 de 2017.

que respondan de manera más oportuna a las necesidades de una sociedad que se perfila para una era de postconflicto.

Palabras clave: Evaluación realista, políticas públicas, teoría del cambio del programa, agencia humana

Clasificaciones JEL: B40, Y20, Y80, Z18.

ABSTRACT

An Introduction to Realist Evaluation and Its Methods: What Works, for Whom, in What Respects, to What Extent, in What Contexts, and How?

This article introduces Realist Evaluation (RE) for public policy assessment. RE finds its justification in the limitations of dominant practices, as impact evaluation techniques, to explain why did an intervention or program exhibit effects, or not. The notion of change or transformation in RE focusses in the study of human agency in context, or the way in which individuals and human collectives act upon private or public interventions. The text delves in methodological proposals for the implementation of a RE, tasks that will entail interdisciplinary efforts to refine alternative theories on what works, for whom, in what respects, to what extent, in what contexts, and how. The introduction of the document includes a brief discussion about the potential of these types of frameworks to contribute to strengthening the design and the implementation of policy initiatives that better respond to the needs of a society that prepares itself for a post-conflict era.

Key words: Realist evaluation, public policy, program theory of change, human agency

JEL **Classifications:** B40, Y20, Y80, Z18

I. INTRODUCCIÓN

Nada es tan práctico como una buena teoría.
Pawson, 2003.

En la era de las políticas públicas basadas en la evidencia, los voceros de este paradigma aseguran que contar con información objetiva y sistemática sobre lo que funciona (y sobre lo que no) permite tomar decisiones sobre el uso de recursos públicos orientadas por criterios de eficiencia y oportunidad. El mantra de las técnicas de evaluación de impacto (TEI) dictaría, por tanto, algo como lo siguiente: en un contexto donde la ciudadanía y las instituciones del Estado exigen a sus gobernantes el rendir cuentas frente al gasto público, "la evaluación de impacto puede ofrecer evidencias sólidas y creíbles del desempeño y, lo que es fundamental, puede determinar si un programa ha logrado los resultados deseados." (Gertler, *et al.*, 2011, p. 4).

La última referencia es extraída de un conocido manual para la evaluación de impacto en la práctica. Un aspecto fundamental en la construcción de esta narrativa es el énfasis que se da a la noción de experimentación (*i.e.* lo experimental, lo cuasi-experimental) para sustentar la tarea del evaluador. En este entendimiento del proceder científico la tarea del experimentador consiste en manipular las condiciones de entorno para examinar si una intervención T tiene un efecto causal sobre un resultado O (Shadish, *et al.*, 2002). En la práctica, los evaluadores de políticas públicas se han apropiado de esta lógica para sustentar estrategias de aleatorización de grupos de tratamiento y control, de modo que sea posible examinar si diferentes programas de educación, salud o de acceso al micro crédito, entre otros, tienen una incidencia directa sobre el aumento de la cobertura escolar, el acceso a medicamentos o la inclusión financiera. El principio metodológico básico es el siguiente:

> Si se implementa de una manera correcta, la asignación genera dos o más grupos o unidades que son probabilísticamente similares en el promedio. Por tanto, cualquier diferencia en resultados que se observe entre esos grupos al final de un estudio son posiblemente debido al tratamiento y no a diferencias entre los grupos que existían al iniciar el estudio (Shadish, *et al.*, 2002, p. 13).

Es oportuno hacer una breve pausa para reflexionar sobre los alcances de este tipo de razonamiento. Resulta pertinente, en particular, preguntarse sobre el tipo

de expectativa (o preguntas) del *policy-maker* y el grado en que estas se traslapan con lo que puede ofrecerle el evaluador-experimentador. Frente a ello, el reconocido manual citado es explícito en advertir que "las evaluaciones de impacto se preocupan por saber cuál es el impacto (o efecto causal) de un programa sobre un resultado de interés. Solo interesa el impacto del programa: el efecto directo que tiene en los resultados" (Gertler, *et al.*, 2011, p. 7). Lo anterior, tal y como lo reconocen sus autores, responde solo a un tipo de interrogante: ¿cuál fue el efecto de T? Más adelante en el texto advierten, sin embargo, que "el trabajo cualitativo puede contribuir a que los responsables de políticas comprendan lo que está ocurriendo en el programa" (*Ibíd.*, p. 17).

El presente artículo se sustenta sobre la premisa de que la respuesta anterior es insuficiente para cumplir la promesa de las TEI de orientar "decisiones sobre políticas" (Gertler, *et al.*, 2011, p. 3). Sin una buena teoría del porqué falla (o no) un programa, las TEI limitan su capacidad de servir como base para hacer recomendaciones a quienes diseñan y ejecutan programas gubernamentales.[1] Se presenta, por tanto, una introducción a la Evaluación Realista (er) como alternativa para responder a algunos interrogantes como ¿Qué funciona? ¿Para quién? ¿En qué aspectos? ¿Hasta qué punto? ¿En qué contextos? y ¿Cómo? Esta escuela metodológica nace en el trabajo de Pawson y Tilley (1997 y 2001), quienes argumentan que todo programa o política implica, necesariamente, la creación de un sistema social (es decir, un compendio de intereses, creencias, incentivos institucionales, recursos, etc.) sobre el cual actúan seres humanos. Visto así, el estudio del cambio social se centra en explorar la forma en que individuos y colectividades expresan su agencia humana en su intento por transformar (o no) su entorno social. Lo anterior implica dar paso a una forma particular de razonamiento que inicia por preguntarse qué condiciones tienen que cumplirse para que los beneficiarios de un programa T obtengan un resultado O en un contexto C.

[1] Deaton (2010), premio nobel de economía en 2015, hace una crítica a las metodologías experimentales y cuasi experimentales (ej. el uso de variables instrumentales) que va en línea con los argumentos de este artículo. Vale la pena citar textualmente un extracto de su texto: "[B]ajo circunstancias ideales, las evaluaciones aleatorias de los proyectos son útiles para obtener una estimación convincente del efecto promedio de un programa o proyecto. El precio de este éxito es un enfoque demasiado estrecho y demasiado local para decirnos 'lo que funciona' en [políticas de] desarrollo, [y] para diseñar políticas, o para avanzar en el conocimiento sobre los procesos de desarrollo. Es improbable que las evaluaciones de proyectos, ya sea mediante ensayos controlados aleatorios o métodos no experimentales, revelen los secretos del desarrollo y, a menos que se guíen por una teoría que esté en sí misma abierta a revisión, es improbable que sean la base de un programa de investigación acumulativo que podría conducir a una mejor comprensión del desarrollo (p. 426).

Dichos preceptos epistemológicos generan un vínculo relevante entre la ER y el presente volumen de *Economía & Región* orientado a la reflexión sobre aportes para la Construcción de Paz desde los Territorios en Colombia. Además de la aplicabilidad en dicho debate de las críticas generales a las TEI esbozadas en el párrafo interior —y que serán profundizadas en el texto— el uso de técnicas experimentales en contextos con poblaciones afectadas por la violencia supone un reto práctico al control de los parámetros técnicos de una evaluación (primariamente) estadística (Bush y Duggan, 2013).[2] A su vez, desde una perspectiva ética, el uso de métodos impersonales para levantar y analizar información (ej. una encuestra estructurada que da poco margen a entablar un diálogo con miembros de una comunidad) puede no solo incidir en la dinámica misma del (post)conflicto (Bush y Duggan, 2013; Bozzoli, *et al.*, 2013), sino también afectar la legitimidad y la apropiación de las prescripciones que emerjan de análisis de políticas públicas (Hesse-Biber, 2013).[3] Para ser justos, debido a la juventud relativa (frente a otros enfoques) de la ER es díficil establecer, al menos de forma empirica, el grado de éxito con el que esta ha efectivamente contribuido a resolver muchas de estas dificultades. No obstante, el esfuerzo y el interés de sus exponentes por responder explicitamente a las debilidades de sus alternativas (Parada, 2007) la convierte en una ruta metodológica que merece ser explorada y discutida, en medio de la búsqueda por fortalecer el diseño y la implementación de iniciativas de política que respondan de manera más oportuna a las necesidades de una sociedad que se perfila para una era de postconflicto.[4]

[2] Al hablar de técnicas experimentales se hace referencia a las TEI en específico, en tanto estas cuentan con una orientación clara frente a la evaluación *causal* de intervenciones sociales. Existen otras técnicas de evaluación con mayor vocación cualitativa o participativa (ej. La Evaluación Constructivista-Responsiva de Stake (2004) o el enfoque hermenéutico de Guba y Lincoln (1989)). Estas se sustenan, sin embargo, en ontologías relativistas del mundo que, por definición, descartan la posibilidad filosófica de la existencia de relaciones causa-efecto. Ello debe suscitar debates sobre la conveniencia de estas últimas para informar estudios causales, como la evaluación de una política pública. Para mayores elementos en esta discusión consultar Parra (2017).

[3] Bozzoli, *et al.* (2013) advierten, por ejemplo, que los esquemas tratamiento-control en este tipo de escenarios pueden deshatar *resentimiento*, en tanto aquellos que se encuentran en el grupo de control pueden sentir que son tratados de una manera injusta. La conducta de estos segunos podría afectar la evaluación y el despliegue mismo del programa analizado. Bush y Duggan (2013), por su parte, hacen hincapié en el riesgo metodológico que conlleva el no centrar el análisis en los actores clave (*stakeholders*, según la nomenclatura utilizada). Gran parte del trabajo de una evaluación experimental, como el diseño de la muestra, se hace previo a la aplicación de formularios de encuesta. Se abre el interrogante sobre la capacidad de tiene un investigador/consultor de identificar relaciones de poder fuera del contexto de la evaluación. Menospreciar las complejidades de las interacciones humanas en contextos específicos, implica un riesgo de hacer una evaluación espurea (Milani, 2009).

[4] El comentario sobre la relevancia (explícita) a los retos del postconflicto se limita a los mensajes esbozados en este párrafo, el cual permite entrever problemas concretos del paradigma experimental en contextos de alta conflictividad social. La invitación al lector es a explorar los preceptos metodológicos de la ER a fondo para per-

El documento está dividido en cinco secciones, incluida esta introducción. La segunda se centra en una crítica sucinta a la lógica básica que guía a los evaluadores-experimentadores, y su lógica de cambio social, con el objetivo de justificar la necesidad de un paradigma de evaluación distinto. La siguiente sección presenta los principios metodológicos de la ER a luz de posibles interrogantes que hace un hacedor de política pública. La cuarta sección se adentra en el campo de la implementación, presentando una ruta tentativa para la ejecución de un proyecto de evaluación. Esta incluye comentarios introductorios sobre criterios de selección de métodos, técnicas de muestreo y análisis de información. El texto finaliza con una reflexión sobre la importancia de repensar la evaluación y la forma en que diferentes actores sociales se benefician de la misma. En últimas, y contrario a cualquier expectativa de generar exclusión epistemológica (contra un método o una disciplina en particular), la ER abre una oportunidad palpable y pragmática para un verdadero diálogo interdisciplinario en el debate sobre la eficiencia y la efectividad de las políticas públicas.

II. ¿QUÉ HACEMOS Y POR QUÉ ES INSUFICIENTE?

Hay políticas e intervenciones que logran mostrar resultados esperados. Pero otras no. Al respecto, Pawson y Tilley (1997) tienden a ser escépticos sobre el control que tiene el hacedor de política sobre el cumplimiento de objetivos específicos de una intervención, incluso a pesar de que la evidencia en la que se sustentan muchos programas públicos que provienen de evaluaciones experimentales. En medio de su discusión citan un influyente estudio de Robert Martinson de 1974, que se propuso examinar todos los reportes publicados en el idioma inglés entre 1945 y 1967 sobre iniciativas para facilitar la rehabilitación de prisioneros sindicados por crímenes en los Estados Unidos. La acumulación de evidencia supondría la generación de un acervo de conocimiento útil para sustentar futuras intervenciones en penitenciarias. La conclusión del análisis, sin embargo, exaltaba lo contrario: aunque algunas acciones estatales habían dado resultado, la realidad era que la vasta mayoría de ellas simplemente no había funcionado.

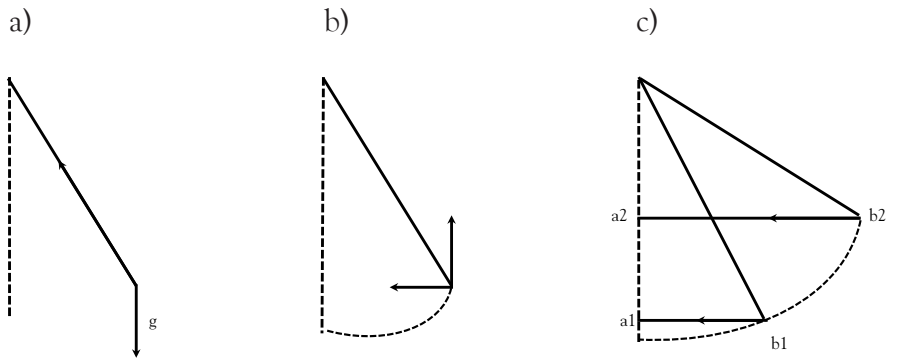
catarse de su potencial para responder a problemas generales de las alternativas de evaluación disponibles, y en particular en contextos como el latinoamericano. Es relevante también precisar que, debido al tono introductorio de este artículo, y el interés que se expresa por dar a conocer una innovación metodológica, no es objetivo del mismo entablar una reflexión crítica sobre la ER. Esta es una tarea que podrá beneficiarse de la puesta en marcha de evaluaciones realistas en campo.

Un ejemplo más contemporáneo de esta (posible) paradoja es el estudio comparado de Glewwe, *et al.* (2013), que evalúa el resultado de dos décadas (1990 – 2010) de políticas educativas orientadas a mejorar el desempeño escolar en países de ingresos bajos y medios. La conclusión, en este caso, prácticamente reproduce el hallazgo de Martinson. Según los autores existe poco soporte empírico para una amplia variedad de características de colegios y docentes, que algunos observadores podrían ver como prioritarias para el gasto en colegios (p. 49). Esta idea luego la complementan al señalar que parte de la ambigüedad proviene de efectos de tratamiento heterogéneos, dónde el impacto de varios insumos depende de circunstancias locales, demandas y capacidades. Esta lectura del problema hace resonancia en el reciente comentario de una académica latinoamericana sobre el por qué, en medio de la insistencia del gobierno de su país en la misma vieja fórmula para promover la calidad educativa, los funcionarios del Ministerio de Educación se sorprendan de que después de años de estos esfuerzos, las escuelas permanecen iguales (Montoya-Vargas, 2014, p. 139).

Aprender supone primero entender.[5] Sin embargo, tal y como lo enuncian las técnicas experimentales, la intención de las mismas consiste en cuantificar y medir efectos. La precaución con la que se presenta la necesidad de utilizar métodos alternativos para estudiar el por qué (Gertler, *et al.*, 2011) hace que la situación de desconocimiento general sobre los mecanismos que permiten el éxito de diferentes políticas públicas no sea del todo sorpresiva. El Gráfico 1 es una réplica del esquema del experimento realista esbozado por Pawson y Tilley (1997), y representa una forma distinta (a la de la prueba y error implícita en las TEI) de investigar las causas detrás de la ocurrencia de un fenómeno empírico (*i.e.* las malas notas en exámenes, el resultado de una elección política). La intuición general, tal y como será discutido a continuación, dicta que toda actividad de generación de conocimiento (el por qué pasa algo) inicia, necesariamente, en el planteamiento de una hipótesis, o una teoría-parcial, del por qué, bajo ciertas circunstancias específicas, es posible observar ciertas regularidades empíricas. Esta preposición, valga resaltar-

[5] Alguien podría aducir, no obstante, que una mascota aprende por repetición a sentarse, saludar y recoger una rama de un árbol. Este es, sin embargo, un tipo de aprendizaje efectivo en un entorno donde los contextos (la relación perro-amo) son relativamente estables. Sin embargo, la investigación en psicología del aprendizaje de Vygotsky (1997) advierte que la adquisición de nuevos conocimientos supone procesos previos de aprendizaje. Ello permite argumentar que, en medio de contextos cambiantes (*i.e.* la heterogeneidad de un país o sociedad), aprender supone entender en qué contextos algunas cosas pueden funcionar mejor que otras.

GRÁFICO 1
El experimento realista



Fuente: Pawson y Tilley (1997).

lo, desafía lógicas de descubrimiento (científico), cuyo objetivo final es predecir un resultado empírico (para una discusión más detallada ver Parra (2016 y 2017)).

La historia que inspira el Gráfico 1, según la presentan sus autores, viene de los registros de Alexandre Koyré sobre los retos que enfrentaron los primeros científicos que intentaron dar respuesta a la pregunta de cómo funciona un péndulo, y la forma en que se empieza a solucionar el interrogante. El primero es el caso de Giovanni B. Riccioli y su equipo de monjes jesuitas que en el siglo XVII, quienes se propusieron, por medio de un esquema de ensayo y error, a descifrar las leyes del movimiento pendular. La explicación general ofrecida en ese entonces se basaba en el modelo de Galileo, según el cual la fuerza de gravedad (un parámetro constante) que actúa sobre la pesa es contrarrestada por la fuerza de resistencia de la cuerda. Para explicar la relación entre el tiempo de oscilación y la longitud del péndulo el experimentador indicó a cada monje que, luego de contar cierto número de oscilaciones, le dieran un pequeño empujón a la pesa para evitar su desaceleración. La tarea se repetiría con diferentes longitudes (experimento) al tiempo que se utilizaba un reloj de agua (tecnología de la época) para cronometrar el tiempo en que el péndulo completase un número determinado de oscilaciones. Narran Pawson y Tilley (1997) qué, no obstante, y pese a que este grupo de religiosos había sido seleccionado por su gran oído musical, los resultados de Riccioli no pudieron ser replicados.

Años después, en 1659, el astrónomo holandés Christiaan Huygens replantea el problema. Según Pawson y Tilley (1997), en lugar de intentar descubrir el resul-

tado de la relación entre longitud y el periodo de oscilación, Huygens inició por generar un modelo sobre los mecanismos subyacentes de la moción del péndulo. El razonamiento matemático señalaba que el movimiento pendular a través de un arco circular no es uniforme. De vuelta al Gráfico 1, el esquema (a) muestra, en primera instancia, la fuerza de la gravedad (g) actuando sobre la pesa. El esquema (b) representa, por su parte, la fuerza de retención de la cuerda y que hala la pesa hacia adentro del semicírculo. Lo que busca representar el esquema (c) son dos momentos particulares de la oscilación, uno de una amplitud menor (a1, b1) y otro de una amplitud mayor (a2, b2), y la forma distintita en cada uno de estos momentos se relaciona con la fuerza interior que hala el péndulo.[6] Con esta teoría en mano, señalan los autores, se abrían nuevas posibilidades experimentales, orientadas a producir algunas condiciones artificiales (contextos C), de modo que diferentes combinaciones de fuerzas (mecanismos M) produjesen un movimiento isocrónico (el resultado O esperado). Y, de hecho, los registros históricos muestran que hubo cientos de variaciones de parámetros, todos impulsados por el dominio incremental de los mecanismos y los contextos de la moción pendular (Pawson y Tilley, 1997, p. 61).

Las enseñanzas que se pueden extraen de este y otros ejemplos (ver Parra, 2016) merecen, sin embargo, algunos matices y aclaraciones. De un lado, es importante resaltar que el uso del modelo matemático de Huygens no tiene un fin deductivo, como el establecer unas propiedades axiomáticas de una función de utilidad para derivar (o pronosticar) un resultado concreto (por ejemplo, el incremento del consumo de X). Este cumple el papel, por el contrario, de generar hipótesis, sujetas a refinamientos, sobre la existencia de mecanismos subyacentes operando en diferentes posibles escenarios. Es por ello que Pawson y Tilley (1997) no hablan de una solución inmediata, sino de la apertura a nuevos experimentos y en diferentes contextos, cada uno alimentado por el conocimiento previamente adquirido sobre las leyes de movimiento del péndulo. Una segunda consideración relevante gira en torno al objeto de análisis de una intervención en escenarios sociales complejos y sobre seres humanos reflexivos (Parra, 2015, 2016 y 2017). Lo anterior implica que los procesos para investigar la sociedad son necesariamente

[6] En caso de amplitudes menores de oscilación, la fuerza interna (a1, b1) es aproximadamente igual al desplazamiento (c, b1) sobre el arco (ver Gráfico 1). Este no es el caso para amplitudes mayores, donde la fuerza interna (a2, b2) es claramente diferente al desplazamiento (c, b2).

distintos a los que seguiría un científico en su laboratorio. Ello no descarta, sin embargo, que puedan aplicarse los mismos principios fundamentales de investigación (Bhaskar, 1998).[7] El resto del texto está orientado por esta premisa y como tal se centra en presentar aspectos metodológicos para poner en práctica una lógica realista de evaluación.

III. REPENSANDO LA NOCIÓN CAUSA-EFECTO

En la introducción de este documento se mencionó el mantra de las técnicas experimentales de evaluación. En el caso de la ER se pueden utilizar también eslóganes para enunciar sus principios. Por ejemplo, toda acción tiene un efecto, la pregunta es para quién y en qué contexto, o, toda hipótesis se alimenta de conocimiento previo (si se quiere, de un prejuicio inicial). La filosofía trascendental realista que sustenta la ER dicta que, dada la brecha (ontológica) que existe entre la percepción y la realidad, una ruta conducente al conocimiento consiste en preguntarse sobre las condiciones que se deben cumplir para que exista X, y no Y.[8] Las primeras pistas para dar una respuesta se encuentran encriptadas en el trabajo de otros investigadores, en la experiencia previa del investigador o en observaciones

[7] Según Bhaskar (1998), en ausencia de acontecimientos espontáneos y ante la imposibilidad de crear sistemas cerrados (por ejemplo, un laboratorio científico) de forma artificial, las ciencias humanas deben enfrentarse al problema del estudio científico directo de fenómenos que solo se manifiestan en sistemas abiertos (por ejemplo, un ambiente donde no es posible controlar todas las condiciones para que la acción T se relacione indeterminadamente con el resultado O). En particular, se deduce de esta condición que los criterios para la valoración y el desarrollo teórico en ciencias sociales, en las que no existen (en principio) situaciones decisivas de verificación, no pueden ser predictivos y deben ser exclusivamente explicativos.

[8] Según Porter (2015) lo que Bhaskar está diciendo es que es posible usar la lógica para establecer cuál debe ser el caso para las características del mundo que observamos sean posibles. Siguiendo a Kant, este es un proceso *trascendental*. Para Bhaskar, cuando se hace una pregunta trascendental, si las premisas son ciertas, y si el camino lógico utilizado es conciso, la respuesta final es confiable. Sin embargo, en cuanto mayor complejidad hay detrás de un fenómeno, menos seguro se puede estar de que el uso de la lógica pura conducirá a respuestas satisfactorias. La objeción se vuelve incluso más fuerte cuando se aplica a acciones de los seres humanos donde el movimiento trascendental solo puede descubrir condiciones necesarias, pero no suficientes, para la ocurrencia de eventos. Vale la pena reseñar de manera muy breve que el artículo de Porter (2015) consiste en una crítica a Pawson a partir del pensamiento de Bhaskar. Sin embargo, el autor concluye que los dos enfoques tienen más coincidencias que diferencias, advirtiendo, no obstante, el riesgo que el argumento transcendental (que en sí da paso a algún grado de especulación, reflexión y verificación) se vea simplificado a un proceso mecánico al servicio del *instrumentalismo burocrático*. El reto del evaluador consiste, por tanto, en no caer en la trampa de simplemente estandarizarlo todo.

preliminares sobre el contexto en que ocurre el evento. De ahí que el *sine qua non* de la política basada en evidencia es un cuerpo de conocimiento cumulativo y progresivo (Pawson, 2006, p. 14), que al conjugar los elementos discutidos en la seccion anterior, da sustento a la siguiente premisa metodológica:

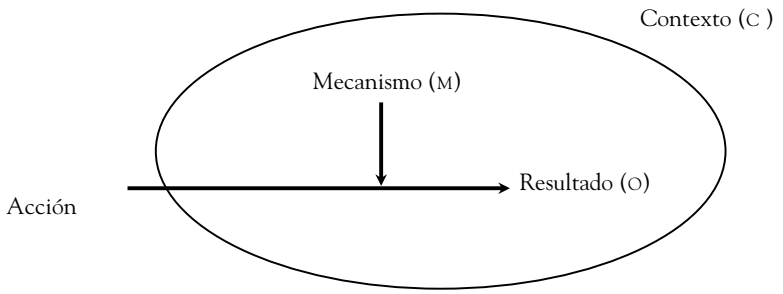
> La taréa básica de la investigación social es el explicar regularidades (R) interesantes, intrigantes, socialmente relevantes. La explicación [inicia] con la sugerencia de algunos mecanismos subyacentes (M) que generan la regularidad y, por tanto, consiste en una serie de preposiciones sobre cómo la relación entre estructura [contextos materiales, culturales] y agencia [tomas de decisiones individuales o grupales] han generado dicha regularidad. En la investigación realista se incluye también la investigación de cómo el trabajo de esos mecanismos es contingente [o, puede variar] y condicional, y, por tanto, es activado en determinados contextos (C) locales, históricos o institucionales (*Ibíd.*, p. 71).

El Gráfico 2 resume la lógica anterior. Su lectura va, por tanto, en la misma línea. Es decir, toda intervención (T), operando en un contexto (C), puede activar un mecanismo (M), produciendo un resultado (O).[9] Un ejemplo puede ilustrarlo mejor. Según Pawson (2013), hace unas tres décadas se introdujeron las cámaras de vigilancia (CCVT) para reducir índices de hurtos en parqueaderos. Dicha iniciativa (T) mostró en sus inicios algunos resultados satisfactorios en Inglaterra (O), en parte porque para los ladrones era una gran novedad y, por tanto, se exponían fácilmente a ser filmados. En la actualidad existe en el Reino Unido cerca de una cámara por cada 32 habitantes, pero también una mayor capacidad de los ladrones de esconderse de las mismas. En el contexto de una emergente sociedad de la vigilancia (C), sostiene Pawson (2013), cualquier persona que entra a un estacionamiento asume que existe un circuito de cámaras.[10] El punto por resaltar

[9] Para claridad conceptual se intenta mantener siempre la misma nomenclatura para hacer referencia a programas (T, como en tratamiento), Mecanismos (M), Contextos (C) y resultados (O). Estas concuerdan con las abreviaciones originales en el idioma inglés y se mantienen así para evitar confusiones.

[10] Un fenómeno emergente, en términos realistas, hace alusión a la forma misma en que surgen y se transforman (o reproducen) eventos en el mundo. Los objetos sociales son producto de la conjugación no lineal entre sus componentes, que, a su vez, son configuraciones de otros subcomponentes, y así sucesivamente. Son resultados no lineales, porque, por ejemplo, así como el agua no tiene las mismas propiedades químicas del hidrogeno o el oxígeno, un colegio con bajo desempeño escolar puede tener buenos y malos profesores, buenos y malos alumnos y buenas o malas instalaciones (Parra, 2015, 2016 y 2017). En este caso concreto se busca resaltar –como se irá discutiendo en el texto– la forma como las circunstancias mismas en las que acontece un

GRÁFICO 2
Causalidad generadora



Fuente: Pawson y Tilley (1997) [adaptado].

es que sin un previo conocimiento sobre el cambio de circunstancias sociales en las que opera un programa de reducción del hurto a la propiedad privada, no es posible estudiar su impacto de una manera rigurosa.

La lógica realista invita a los evaluadores a repensar, por tanto, la noción de causa-efecto que engloba su actividad evaluadora. En los esquemas experimentales, cuyo artefacto metodológico para responder preguntas sobre el efecto es la modelación estadística (*i.e.* la econometría), prevalece una noción *sucesionista* del cambio social, donde se asume que O siempre es precedido por T (Parra, 2016). En dicho caso, el investigador asume que la relación causal entre T y O es externa o exógena, y por tanto su labor consiste en "aplicar las medidas y los controles de la manera más sistemática y rigurosa posible [...] esperando [observar] alguna diferencia neta en los resultados entre sujetos [tratados y no tratados]" (Pawson y Tilley, 1997, p. 33). Según las discusiones tratadas en la sección anterior, esto se traduce en una práctica de evaluación problemática. La ER se adhiere, por el contrario, a un tipo de causalidad generadora, enfocada en estudiar la transformación potencial de un fenómeno (*Ibíd.*, p. 34) para producir un resultado concreto en un contexto específico.

programa o política social son fenómenos sociales que pueden cambiar y transformarse como resultado mismo de una intervención. Se incluyen acá valores y expectativas que se forman individuos y grupos frente a distintos escenarios sociales. En economía del comportamiento se habla de un fenómeno algo similar, bajo el rótulo de preferencias endógenas (Parra, 2013).

Así las cosas, dentro de la nomenclatura de la ER es importante reflexionar en torno a tres conceptos que ya se han venido incorporando en el texto, y que permiten poner a operar una lógica causal generadora: mecanismos, contextos y resultados. El último de ellos es el que requiere de menos aclaraciones, en tanto representa una noción universal y recurrente de todos los esquemas de evaluación. Un resultado representa un evento (por ejemplo, el cambio en el desempeño escolar, la evolución de un indicador de cobertura en salud) esperado (o no) de una intervención (T). Sin embargo, es en la noción de los dos conceptos restantes donde la ER toma distancia de otras propuestas metodológicas. Al respecto, van Belle, *et al.* (2016) proponen la siguiente distinción frente a la forma en que estos se entienden y se entrelazan en una lógica experimental (causalidad sucesionista) y aquella asociada con un esquema realista (causalidad generadora). En el primer caso:

> [Se asume la existencia] de patrones causales frecuentes que se desencadenan bajo condiciones generalmente desconocidas y con consecuencias indeterminadas. Un mecanismo explica al abrir la caja negra, develando los engranajes y las ruedas de la maquinaria interna. Este proporciona la cadena continua y contigua de vínculos causales o intencionales entre los [factores] explicativos y los [factores] explicados (van Belle, *et al.*, 2016, p. 3).

En un esquema realista, por su parte:

> [Un mecanismo corresponde a] una entidad no observada que, una vez activada, genera un resultado de interés [...] [Por lo tanto] un análisis causal consiste en la identificación de la configuración [específica de factores] que vincula resultados a mecanismos activados por [un] contexto [particular] (*Ibíd.*, p. 3).

Queda implícita, por tanto, una noción sobre la naturaleza de una política o un programa público, y sus posibles efectos, que dista de una visión mecanicista más propia de lógicas experimentales de evaluación. Al vincular mecanismos con las propiedades específicas de un contexto físico (como un barrio o una ciudad) o espacial (como la era de consumo o el tiempo transcurrido tras una reforma de salud) se asume que los efectos de toda intervención en un sistema humano son mediados (o activados) por la acción (u omisión) de los actores (individuos, colectividades) que hacen parte del mismo (Pawson, 2013). Puesto en términos (algo) más formales, las intervenciones no funcionan por sí mismas; estas solo tienen

efecto a través del razonamiento y las reacciones de sus receptores (Pawson, *et al.*, 2011a, p. 519) cuya agencia (o comportamiento) dependerá de configuraciones sociales que se encuentran inmersas en especificidades históricas y culturales.

Un corolario de lo anterior consiste en tipificar los contextos no como escenarios inertes o neutrales donde opera una política o programa público —como se asume en escenarios de tratamiento y control— sino como las propiedades de espacios humanos, geográficos o institucionales (una prisión, un colegio, un barrio o comunidad) presentes en un compendio preexistente de reglas sociales, normas, valores e interrelaciones (Pawson y Tilley, 1997, p. 70). De ello se desprende, como se exalta a continuación, la necesidad de integrar el estudio de las propiedades de ese contexto específico, y teorizar sobre como estas condicionan (mas no determinan) ciertos comportamientos en individuos o colectividades. Este último es un componente básico (y no secundario o complementario) de cualquier proyecto de evaluación de una intervención pública o privada encaminada a transformar las condiciones (por ejemplo, pobreza o exclusión) en un sistema o subsistema social.

IV. LA ER EN LA PRÁCTICA: ALGUNOS ASPECTOS METODOLÓGICOS

Los supuestos inmersos en diferentes lógicas de evaluación se ven plasmados en la selección de metodologías y métodos de análisis de información. Quienes se encuentran familiarizados con las técnicas de evaluación de impacto han estado expuestos a argumentos sobre los beneficios, en términos de trazabilidad y replicabilidad, de las herramientas cuantitativas. Dentro de este primer paradigma la tarea del evaluador consistirá, por tanto, en implementar técnicas de análisis de correlación controlada que son utilizadas para verificar los mecanismos que se considera que están mediando (o activando) el efecto observado (van Belle, *et al.*, 2016). El uso de la teoría (principalmente económica) en este caso se reduce, por tanto, a expresar principios o axiomas (ej. la racionalidad de los consumidores y productores) que permiten deducir algunas generalidades a partir de observaciones empíricas.

De las ideas discutidas hasta el momento es posible inferir una práctica distinta en la tradición de la ER. Apelando nuevamente al lenguaje de Pawson y Tilley (1997), dado que los programas funcionan a partir de la introducción de nuevas

ideas o recursos en contextos que contienen relaciones sociales preexistentes, la tarea crucial de la evaluación es incluir (a través de la generación de hipótesis y el planteamiento del diseño de investigación) evidencia sobre el alcance de estas estructuras preexistentes para habilitar o no el mecanismo intencionado de cambio. Esta teoría de cambio, como se discute en los apartes siguientes del texto, constituye el corazón de la ER, la cual se pone, como gran objetivo, iniciar un proceso iterativo de refinar dicha teoría por medio de la combinación de diferentes herramientas de análisis de información cuantitativa y cualitativa.

A. La teoría del programa (M + C = O)

La noción (general) de una teoría de cambio de un programa no es un concepto exclusivo de la ER. En el manual de evaluación de impacto en la práctica publicado por el Banco Mundial, la Teoría de Cambio (TC) se encuentra explícitamente definida 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, p. 22). No obstante, en este momento es claro el sentido causal sucesionista que inspira la misma. Mientras la TC se basa en una versión rígida sobre el deber ser de la operación de una intervención T, una Teoría del Programa, como ha sido denominada en un paradigma de evaluación realista, se plantea más a modo de una pregunta-teoría (suele utilizarse el término teoría intermedia), sujeta a ser refinada durante el proceso de evaluación (Blamey y Mackenzie, 2007). Un interrogante genérico sonaría, por ende, de la siguiente manera:

> ¿Cuáles son las condiciones sociales y culturales necesarias para que mecanismos cambiantes entren en operación y como se encuentran distribuidas en diferentes contextos del programa? (Pawson y Tilley, 1997, p. 77).

La anotación formal de la Teoría del Programa en la ER viene representada por una simple operación aritmética denominada una configuración CMO: C + M = O. Explica Pawson (2006) que, si bien se trata de una terminología torpe, esta presenta un contraste marcado con la visión sucesionista —típica de una TC– que prioriza la búsqueda de regularidades en resultados. Pero más relevante aun, vale resaltar, es la forma en que la lógica CMO, al contemplar explícitamente la interacción entre contextos y mecanismos en la explicación de un resultado, materializa

la pregunta básica sobre el qué funciona, para quién, en qué aspectos, hasta qué punto, en qué contexto y cómo.

Es momento de traer a colación ejemplos de configuraciones CMO al servicio de evaluaciones de intervenciones reales. Un caso ya emblemático, discutido por Pawson (2013) y sus colegas (Pawson, *et al.*, 2011a) es el de la efectividad de la legislación inglesa que prohíbe el consumo de tabaco en vehículos que transportan menores de edad. En palabras de los autores, y en medio de la necesidad de conceptualizar una intervención legislativa como un proceso (una cadena de valor) que debe navegar entre las expectativas de legisladores, de grupos de presión, de las agencias que las aplican y las promueven y de las distintas sensibilidades de públicos fumadores y no fumadores, este tipo de evaluación representa un desenlace con un típico y abundante grupo de certezas e interrogantes (Pawson, 2013, p. 160). A partir de la consulta previa de intervenciones que podrían considerarse relevantes (he aquí la importancia de involucrar expertos con conocimiento del sector en la fase de discusión de diseño de la evaluación), se identificaron una serie de elementos, expresados en preguntas y subpreguntas, que permiten esbozar una primera aproximación a la teoría de cambio del programa (Cuadro 1).

CUADRO 1

Ejemplo de interrogantes preliminares CMO

No.	Pregunta
1	¿Es el problema lo suficientemente grave para justificar una nueva ley? — ¿La exposición al humo por parte de fumadores pasivos incrementa el daño en su salud? — ¿Qué niveles de toxicidad se encuentran en un vehículo cuando se fuma en su interior?
2	¿Es factible encontrar apoyo en el público para este tipo de legislación? — ¿Cuál podría ser el nivel de apoyo entre fumadores? — ¿Qué motivación se encuentra detrás del apoyo de la población?
3	¿Es posible encontrar grupos de presión resistiendo la prohibición? — Existen antecedentes de oposición a este tipo de iniciativas por parte de empresas de tabaco?
4	¿Es institucionalmente posible aplicar una ley de este tipo? — ¿Cuáles pueden ser las mayores barreras/líneas de apoyo institucionales (ej. Legales, procesales) al momento de implementar la ley?

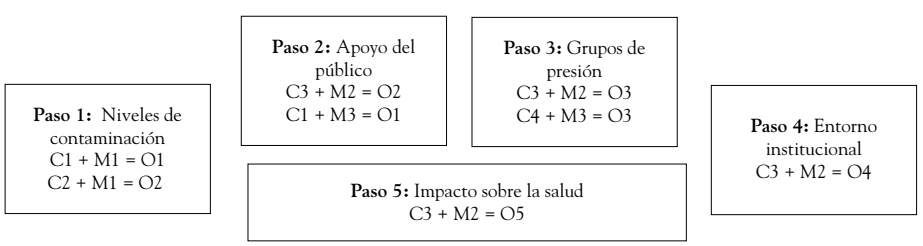
Fuente: Pawson, *et al.* (2011a) [adaptado].

Este primer caso pone sobre la mesa un punto que, si bien podría sonar obvio, es importante aclarar. La teoría de un programa, más en intervenciones con este grado de complejidad, emerge de la conjugación de distintas configuraciones CMO (o sub teorías), como se ilustra en el Gráfico 3. Dicho esquema representa una forma de enlazar las diferentes dimensiones de la política de acuerdo con las preguntas enunciadas en el Cuadro 1 y como tal representa un punto de partida creíble para los fines de una ER. Parafraseando a Pawson, *et al.* (2011a), a fin de poder generar recomendaciones de política basadas en evidencia sobre la factibilidad que la prohibición de fumar en un vehículo con niños funcione es necesario contar, al menos, con información relevante sobre la toxicidad del cigarrillo, el apoyo público de reformas similares, el poder y las estrategias de grupos de presión y sobre el entorno institucional en que sería implementada una ley de este tipo (por ejemplo, la relación entre el poder ejecutivo, intereses en el poder legislativo, etc.).

La lectura del Gráfico 3 permite entrever posibles relaciones teóricas al interior de cada componente y entre los mismos. En este caso podría suponerse, por ejemplo, que el apoyo del público dependerá de la información médica disponible, pero también de qué tan exitosos son los grupos de presión en transmitirla de una manera estratégica. A su vez, los efectos finales sobre la salud serán un resultado (más) directo de la reducción de niveles de contaminación, conjugados con la forma (o los procesos) misma de la implementación. El reto de la ER es, por tanto, refinar estas subteorías (CMO), de modo que se pueda entregar una teoría más robusta (basada en investigación con fuentes secundarias y primarias) al hacedor de política. El siguiente ejemplo presenta con más detalle una configuración CMO correspondiente a uno de los compontes de la teoría de cambio propuesta (Gráfico 3).

GRÁFICO 3

Una Teoría de Cambio para el análisis de la ley antitabaco en el Reino Unido



Fuentes: Pawson, *et al.* (2011a) y Wong y Papoutsi (2016) [adaptado].

> 1) debido al espacio de la cabina de un carro (C), 2) bajo las peores condiciones de ventilación (C), y 3) dado los picos máximos de ventilación (C), la evidencia nos permite decir que el fumar en un vehículo genera partículas finas de concentración que son 4) muy raramente experimentadas en el marco de los estudios de calidad del aire (M), y que por tanto constituyen un riesgo significativo a la salud (O) (Pawson, *et al.*, 2011b, p. E684).[11]

Se han añadido, a modo de ilustración, iniciales que resaltan la interacción entre contextos y mecanismos para generar resultados particulares. El ejemplo revela un caso concreto en la forma que se puede materializar la lógica que se ha venido discutiendo –entreviendo, en otras cosas, que la nomenclatura CMO no pone una camisa de fuerza a la cantidad de Cs u Os que el investigador puede adjudicar a una sola O–. La evidencia que se genera en este caso (el estudio de Pawson, *et al.* (2011b) documenta fuentes, cifras e hipótesis intermedias) indica que, si bien el cigarrillo tiene componentes nocivos en sí mismo, contextos particulares como un vehículo con poca ventilación detonan la producción de partículas que aumentan los daños de exposición al mismo.

Otro ejemplo relevante, esta vez de una ER finalizada, es el análisis emprendido por Horrocks y Budd (2015) sobre el programa eGovernment for You (EGOV4U) del gobierno en línea financiado por la Comisión Europea entre 2010 y 2011. La intervención consistió en que los miembros de un consorcio de oficinas municipales de gobierno de cinco países europeos recibieron durante ese período un apoyo financiero en el desarrollo de iniciativas de Tecnologías de la Información (ICT) para generar canales de comunicación y mecanismos para focalizar esfuerzos e iniciativas de política en poblaciones vulnerables. Siguiendo la lógica del qué funciona, para quién, en qué aspectos, hasta qué punto, en qué contexto y cómo, los autores proponen una teoría del programa (con configuraciones CMO) de la siguiente forma:

> La identificación de los servicios electrónicos apropiados y dirigidos con precisión a las necesidades de las personas económicamente/socialmente desfavorecidas/

[11] Las nomenclaturas alusivas a la configuración CMO fueron añadidas por los autores. En adelante, cuando se trate de teorías de programa, se agregan —y en caso de que ya existan, se modifican de acuerdo con la nomenclatura utilizada— letras C, M y O para hacer explícita, en cada configuración, los contextos, mecanismos y resultados plasmados en cada una de ellas.

> excluidas **(C)**, y diseñados y entregados de tal manera que tenían una buena oportunidad de ser eficaces en abordar necesidades identificadas **(O)** –y que eran sostenibles– estaba más allá de la capacidad y el alcance de los socios de EGOV4U que operan solos. Por lo tanto, las [Redes Multi-Canal R-MC] son el mecanismo **(M)** necesario para alcanzar estos objetivos. Adicionalmente, las R-MC fueron cruciales para el desarrollo/mejora de la reputación [de gobiernos y entidades prestadoras de servicios]. La formación de R-MC que incluyeron otras organizaciones e intermediarios actuales o potenciales y los usuarios del servicio electrónico **(M)** se consideró, por tanto, una característica fundamental de los procesos de implementación **(O)** (Horrocks y Budd, 2015, p. 55).

En el texto, vale la pena anotar, los autores insisten en que debido a los argumentos ontológicos del realismo sobre cómo funciona y se transforma una sociedad —apelando a sus mecanismos generadores (Parra, 2016 y 2017)— la transferencia de lecciones de programas e iniciativas no emerge de generalizaciones hechas sobre hallazgos empíricos sino a través de un proceso de construcción teórica (Horrocks y Budd, 2015, p. 53). Se desprende también de su trabajo la importancia de insistir sobre el hecho de cómo el objetivo del evaluador es el poner a prueba diferentes CMO y en diferentes niveles, implicando que la evaluación no gira exclusivamente en torno a un solo gran objetivo O (ej. mejorar la focalización de servicios). Así, por ejemplo, una teoría refinada que se desprende de su análisis indica que EGOV4U creó R-MC conducentes a la generación de espacios de información (páginas web) y formación en uso de medios virtuales para mujeres, personas de tercera edad y veteranos de guerra, reconocidos con premios nacionales de innovación (resultado parcial 1). Esta sinergia (voluntarios de poblaciones vulnerables y escuelas de formación) se expandió geográficamente e incorporó nuevos socios y aliados (resultado parcial 2), y con ello la generación de nuevos hábitos sociales positivos frente a este tipo de iniciativas de apoyo a grupos de poblaciones específicos (resultado parcial 3).

Para finalizar esta sección se traen a colación algunos ejemplos de configuraciones CMO más puntuales, y que ponen en relieve la utilidad de la ER en diferentes campos y sectores económicos y sociales. La primera proviene de un estudio sobre prácticas administrativas en hospitales en Ghana:[12]

[12] Este ejemplo es extraído directamente de la página de *BetterEvaluation.org*. La teoría del programa que presenta es una modificación (en cuanto a las nomenclaturas) de la teoría original publicada en este sitio.

> Un equipo de gestión hospitalaria puede lograr un mayor compromiso organizacional si sus prácticas de gestión actúan sobre [lógicas de intercambio económico y social]. Cuando el personal percibe altos niveles de apoyo a la gestión o apoyo organizativo percibido **(T)**, desarrollarán comportamientos [que denotan grados de compromiso adicionales] como trabajar tarde, [o de ser más solidarios entre si] **(O)**, sobre la base de la reciprocidad **(M)**, incluso en los hospitales con márgenes limitados de libertad en materia de contratación, escalas salariales, promoción y despido **(C)**.

Un segundo ejemplo proviene del trabajo de Westhorp (2013), quien hace parte del equipo de la empresa de consultoría Community Matters con sede en Australia y que trabaja también en alianza con organizaciones como United States Agency for International Development (USAID). La ER en mención se ocupa de revisar efectos de un programa implementado en Estados Unidos para atender a la primera infancia. La teoría (refinada) del programa, en este caso, dicta lo siguiente:

> En contextos de alta pobreza y altas exigencias de tiempo a los padres (por ejemplo, para los trabajadores pobres) **(C)**, y/o en contextos de alta pobreza pero también con alto apoyo social **(C)**, los programas que utilizan un modelo de formación de padres sin negociar el contenido o los métodos con los padres **(C)**, pueden ser experimentados por padres como una carga adicional, lo que aumenta sus niveles de estrés **(M)** y, por ende, probablemente contribuirán al aumento de actitudes negativas hacia las exigencias de la crianza de los hijos **(M)** resultando en una crianza menos sensitiva **(O)** y que es probable que genere (...) resultados menos positivos para el desarrollo del niño **(O)** (Westhorp, 2013, p. 380).

El último caso ilustrativo de un esquema CMO en la práctica es referenciado por Pawson (2013) en su último libro sobre métodos de evaluación realista.[13] Este se centra en un programa de generación de capacidades en una comunidad para atender las necesidades psicomotoras de niños y niñas en edad preescolar. En

[13] Este ejemplo, valga la aclaración, es utilizado por Pawson y Manzano (2012) para entablar una crítica a prácticas que usan postulados realistas sin el rigor necesario en fases de verificación de las configuraciones CMO. Es decir, se señala que los autores del mismo hacen selecciones arbitrarias de extractos de entrevistas que se ajusten a sus configuraciones CMO. El estudio tiene, por tanto, un problema en términos de recolección y análisis de información. Pese a lo anterior, Pawson y Manzano (2012) reconocen que este estudio hace un buen trabajo en generar hipótesis iniciales sobre escenarios del programa y comportamientos individuales. Por tanto, se considera un buen ejemplo de una configuración CMO.

palabras del autor (el cual reinterpreta el estudio original para presentarlo en la nomenclatura realista), una posible configuración sería la siguiente:

> La estrategia de agrupar a niños y niñas con necesidades similares puede funcionar mejor para menores con dificultades lingüísticas y motoras finas **(C)**, al permitir que estos miren y copien a otros para desarrollar sus propias habilidades motoras finas **(M)** y resultado en una mayor disposición para intentar nuevas actividades en el jardín de infancia y el hogar **(O)** (Pawson, 2013, p. 20).

Para resumir, los ejemplos anteriores hacen explícita la forma en la que la ER pone en práctica su intención de analizar un programa a partir de la lectura de un proceso con múltiples configuraciones CMO que interactúan entre sí. Esta es una tarea que, a diferencia de otras lógicas de evaluación, dista de ser lineal, y como tal, debe alimentarse de la consulta de diferentes fuentes y tipos (cualitativa, cuantitativa) de información. La siguiente sección del artículo se centra en discusiones más relacionadas al análisis en sí, y a la forma en que un equipo consultor que se rige por una lógica realista debe poner en marcha un proceso de teorización y refinamiento de teorías del programa para dar respuesta al qué funciona, para quién, en qué aspectos, hasta qué punto, en qué contexto y cómo.

B. Sobre recolección y análisis de información

Un evaluador experimentado sabe que el éxito de una evaluación depende de diferentes factores y decisiones en campo. Una característica de las evaluaciones experimentales es que estas dedican gran parte de dichas discusiones a los aspectos técnicos del proyecto (*i.e.* el muestreo estadístico, las técnicas de aleatorización, etc.). Sin embargo, como lo reconoce Milani (2009, p. 49), las preguntas en torno al método y el diseño no son irrelevantes, pero al parecer, en medio del denominado paradigma de evaluación basado en la evidencia, se ha borrado el papel crucial del comportamiento (político) de los individuos en los programas analizados.

Esta última idea debe matizarse, en tanto parecería restar importancia al diseño de una evaluación. Cabe recordar, sin embargo, que la lógica básica de una ER consiste en proponer una teoría sobre el cómo funciona un programa y refinarla. Como tal, se trata de un proceso iterativo que implicará, más allá de un simple ejercicio de reconocimiento de diferencias estadísticas entre subgrupos, llevar a

cabo un análisis de triangulación y corroboración de relaciones cruzadas entre elementos y componentes de una intervención (Pawson y Manzano, 2012). Por tanto, el diseño hace parte esencial de un esquema realista de evaluación, pero entendiendo el mismo como una fase que debe:

> [...] dar pistas sobre los mecanismos de cambio actuales del programa [...] Dicho conocimiento puede ser extremadamente crítico al analizar la posibilidad de generalizar [el alcance] del programa para tomar decisiones de replicarlo o adaptarlo a nuevos escenarios. Un buen diseño debe dar pistas sobre los contextos y los mecanismos por los cuales funciona el programa (Sridharan y Nakaima, 2011, p. 141)

Este razonamiento ofrece elementos para responder a interrogantes frecuentes entre evaluadores frente a temas como la selección y el tamaño de una muestra, la interpretación de información y el delineamiento de prescripciones de política pública. Una respuesta general a todo ello se desprende de la lógica metodológica esbozada hasta el momento; todo depende de la teoría del programa, y, por tanto, del soporte argumentativo con el que cuente el investigador para defender su teoría (refinada). Así, por ejemplo, un muestreo realista debe ser diseñado para corroborar los elementos relevantes dentro de la teoría del programa y su definición final dependerá, en gran medida, de hallazgos preliminares una vez iniciado el trabajo de campo (Emmel, 2013; Manzano, 2016).

Es momento de hacer referencia a métodos para construir y, sobre todo, refinar configuraciones CMO. Los debates vigentes en ciencias sociales parecen haber llegado al acuerdo de las bondades de los llamados métodos mixtos. Desde la economía, disciplina que se ha posicionado como fundamento metodológico de las evaluaciones experimentales, se ha argumentado, por ejemplo, que, al primero aplicar una encuesta a gran escala, y luego hacer seguimiento con entrevistas a profundidad o grupos focales es posible enriquecer hallazgos cuantitativos (Starr, 2014). Otra razón para combinar información cuantitativa y cualitativa, y tras la cual se intenta equiparar la jerarquía entre ambos enfoques, apela al hecho que un análisis simultáneo puede ayudar a entender resultados inesperados que surgen en cada tipo de recolección de datos (Starr, 2014). No es muy claro, sin embargo, cómo este tipo de argumentos coincide con el lineamiento general de la evaluación de impacto, presente al inicio de este texto, donde un enfoque sirve para medir y el otro para ofrecer ideas de la posible explicación detrás de los resultados de una política o programa.

La noción de causalidad generadora de la ER hace evidente, sin embargo, que muchos de los argumentos a favor de los métodos mixtos de investigación carecen de peso metodológico. La razón más evidente de ello es que los tipos de información cuantitativa y cualitativa (incluyendo la lógica bajo la cual son recolectadas) responden a supuestos y necesidades diferentes. A partir del realismo se aduciría, por ejemplo, que mientras que la estadística es útil para medir o describir regularidades (ej. tendencias en indicadores de cobertura o desempeño), las conversaciones directas con actores políticos y sociales –o la observación directa en campo– cuentan con un mayor potencial para capturar relaciones humanas complejas (Sayer, 2000; Porpora, 2015). Queda por tanto explícito el argumento a favor de la combinación de formas y lógicas de análisis de información, pero también de la mayor jerarquía en importancia de las indagaciones sobre las cualidades (y no las magnitudes o cantidades) de los objetos (ej. programas) analizados. Usando los términos de Patomaki (2003), debido a naturaleza de los procesos causales bajo estudio (ej. generativos y dependientes de su activación por parte de agentes sociales) los métodos y el lenguaje cualitativo son necesarios para identificar las estructuras y los poderes causales relevantes.

Hechas esas aclaraciones, se procede a continuación a exaltar algunas particularidades de los métodos de la ER en la práctica. Quizás el más relevante –y del cual se desprenden enseñanzas extrapolables a otros métodos cualitativos de análisis– es el de la entrevista realista (Pawson, 1996; Smith y Elger, 2014; Manzano, 2016). En su artículo clásico publicado en The British Journal of Sociology, Pawson (1996) advierte que más allá de la distinción binaria entre una entrevista estructurada –con fines deductivos– y una abierta y/o semi-estructurada –más coherente con lógicas de inferencia inductiva–, el objetivo de una entrevista consiste en permitir confirmar o falsar o, sobre todo, refinar una teoría. Las lógicas más clásicas de indagación en campo, señala el autor, tienden a asumir que el objetivo del despliegue de un instrumento (ej. un formulario de preguntas) es el entrevistado y no, como se plantea en la ER, la teoría del programa. Siendo ese el caso, el entrevistador no es un simple captador de información; el objetivo consiste en crear una situación en la cual los postulados/conceptos teóricos bajo investigación estén abiertos a la inspección.

Para ilustrar lo anterior se citan a continuación dos trabajos académicos que hacen reflexiones explícitas sobre el arte de la entrevista realista en la práctica. El primero de ellos es el de Manzano (2011 y 2016) quién centra lo discusión en torno a los resultados de un programa implementado en Inglaterra para enfrentar la

congestión en hospitales públicos. El Acto de Salud Comunitaria de 2003, narra la autora, propuso un esquema de multas a los gobiernos locales que se tomaran más de tres días para trasladar a un paciente declarado de alta por autoridades médicas. Los evaluadores en este caso documentan el éxito alcanzado por lógicas punitivas de este tipo en sociedades del norte de Europa en los años 90, pero indican que no hay claridad sobre las relaciones causales (el por qué y el cómo) encriptadas en las mismas. La ER reportada por Manzano (2011) busca contribuir con una pieza del rompecabezas, y por ello no se acudió a un muestreo probabilístico –el cual, argumentan los evaluadores, no hubiera servido para capturar las complejidades de la intervención– pero sí a uno basado en discusiones con miembros de hospitales y de los equipos de las entidades prestadoras de servicios a fin de visualizar diferentes circunstancias en las que podría tener éxito, o no, tal iniciativa:

> En el caso del programa de altas retrasadas, este esquema se puso a prueba en los años noventa en tres países escandinavos y luego en Inglaterra a principios de los años 2000 en más de 164 fidecomisos [o áreas de atención regional] y sus respectivos departamentos de servicios sociales. Cada una de las teorías del programa transformadas a nivel local representa un ejercicio de prueba a partir del cual se aprenden lecciones [...] Este estudio, localizado en solo un fidecomiso y con datos de sólo 14 [pacientes hospitalarios], se enfrentó al reto de establecer generalizaciones a partir de un pequeño número de estudios de caso, mientras que se evaluaba un programa complejo. No puede considerarse, sin embargo, como un proyecto único porque fue construido desde el aprendizaje de estudios preliminares (...) Por consiguiente, los hallazgos de esta implementación local indicarán algo acerca de la teoría del programa ampliada (las multas funcionan para reducir retrasos en altas hospitalarias de pacientes) como una teoría general del cambio. Así, es posible extraer lecciones de todo el mundo y para el público en general (Manzano, 2011, p. 28).

La primera etapa de la evaluación fue de inspección, y se basó en la recolección y el análisis de información etnográfica (observación) y de 34 entrevistas (13 con pacientes, 12 con trabajadores de hospitales y las restantes como miembros de equipos de servicios sociales). Inspeccionar, en el caso de una ER, implica articular teorías iniciales para identificar cómo las circunstancias de contexto de algunos usuarios/programas pueden llegar a impactar el comportamiento y [su] efectividad (Manzano, 2011, p. 13). En tal medida, la autora, con base en su

experiencia, recomienda hacer preguntas del tipo ¿cómo era su trabajo antes del programa? ¿cree que este programa va a funcionar para todos los implicados? ¿podría especificarme en qué circunstancias y para que personas podría ser más efectiva una intervención de este tipo?, que permiten definir el qué, cómo, para quién y en qué circunstancias. Ejemplos más concretos de esta lógica se ven plasmados en los siguientes interrogantes:

> ¿Qué características de este paciente sugieren que es probable que se demoren en ser dado de alta del hospital? Estoy pensando en edad, estado de salud mental, finanzas [...] ¿Cómo crees que el nuevo sistema de multas ha impactado en cómo el personal de servicios sociales se ocupó de este caso? Estoy pensando qué pueden estar haciendo de manera diferente de lo que solían hacer antes de que se implementara el nuevo programa (Manzano, 2016, p. 353).

En la medida que el equipo evaluador se familiariza con el programa o con la intervención, es posible elaborar preguntas más específicas sobre posibles mecanismos y contextos que interactúan en la generación de resultados observados a lo largo del proceso. Es acá donde conceptos como el de la triangulación de información, que por lo general son presentados de forma abstracta y difusa en los manuales de investigación cualitativa, son útiles para describir la lógica de teorización y refinamiento de teorías y sub-teorías (CMO) durante el proceso mismo de levantamiento de información —y menos, como es común, en la engorrosa tarea de identificar teorías emergentes una vez terminado el proceso de recolección—. Como argumenta Manzano (2016), los significados atribuidos por el evaluador a las respuestas anteriores o las situaciones observadas se discuten a la luz de la divergencia potencial entre ambas y se presentan al entrevistado mientras que se va obteniendo la evidencia, lógica que se materializa al indagar sobre elementos del tipo "usted me dice esto, pero yo observe aquello en mi visita" o "su compañero me contó X, pero veo que su opinión difiero de ello".

> **Evaluador:** ¿Por qué cree que este paciente fue dado de alta a un asilo de ancianos? ¿Cree que podría haber ido a su propia casa en lugar de un hogar de ancianos? Y lo digo porque una de las teorías sobre esta política es que para acelerar las altas hospitalarias debe enviarse a la gente más temprano a los hogares de cuidado. ¿Correcto?
>
> **Enfermera:** Hmmm [...] Creo que, en ese momento, este paciente podría haber ido a casa y haber atendido en casa con un gran paquete de cuidado. Podríamos

> haber organizado tres a cuatro visitas de atención domiciliaria al día, y una visita en la noche. Creo que ha entrado en la atención [en el asilo] porque había estado en el hospital durante mucho tiempo y estaba asustado de estar por su cuenta. Además, tenía algunos problemas médicos que necesitaban monitoreo. Y creo que, posiblemente, también buscaba paz mental (Manzano, 2016, p. 355).

En el extracto anterior, que da cuenta de un diálogo entre un evaluador y una enfermera, se hace evidente la intención del investigador de poner a prueba versiones sobre los mecanismos del programa, y que se pueden obtener en diferentes fuentes institucionales o no institucionales (como por ejemplo parte de la narrativa oficial condensada en documentos de política, un diálogo preliminar con un alto directivo del ente ejecutor e incluso indicadores cuantitativos relevantes sobre la prestación de servicios de salud). Dicho ejercicio de contraste puede llevarse a cabo en entrevistas individuales, y, sobre todo, con actores que se consideran estratégicos en tanto cuentan con información de primera mano sobre el día a día de la actividad que se está examinando (una enfermera, en este caso, es un informante estratégico, más si cuenta con ciertas características, como experiencia y antigüedad en su cargo). En este caso concreto, según aclara la autora, se buscaba explorar el mecanismo que podría explicar la salida de pacientes hacia asilos u hogares especializados de cuidado para acelerar el proceso de descargas hospitalarias.

El segundo ejemplo es el de Jackson y Kolla (2012) el cual discute una ER sobre una intervención para generar capacidades en padres y madres primerizos frente al cuidado de menores con problemas de salud mental. Los administradores de unas instituciones de salud estaban interesados en conocer más sobre el efecto de la figura del padre comunitario (PC) –un individuo sin credenciales médicas o formación certificada en cuidado infantil, pero con experiencia como padre de familia– como actor de soporte a las nuevas familias. Parte de la teoría del programa, informada por la amplia experiencia (en algunos casos, de más de 10 años) de los funcionarios en la implementación de políticas de cuidado infantil, indicaba que este tipo de apoyos era esencial en comunidades de alta vulnerabilidad (pobreza, desempleo, informalidad laboral, escasa protección social). Para probar diferentes configuraciones CMO se llevaron a cabo once entrevistas con PC, educadores miembros del programa y enfermeras de instituciones públicas vinculadas o no al programa. En palabras de las autoras, los participantes fueron elegidos con base en su familiaridad con estos programas y el grado de profundidad de su experiencia en los mismos para proporcionar información sobre el funcionamiento en una variedad de contextos (Jackson y Kolla, 2012, p. 342).

Una consigna que se desprende de las lógicas de investigación en el paradigma realista —y que ponen en manifiesto las complejidades y las múltiples situaciones contingentes que surgen en la investigación social (Parra, 2016)— es que es importante explorar diferentes formas de análisis e interpretación según el tipo o las fuentes de información al servicio del investigador (Pawson y Tilley, 2001). La propuesta de Jackon y Kolla (2012) representa, por tanto, solo una posibilidad de análisis (y, por ende, no debe ser tomada como una camisa de fuerza). Las autoras, en este caso, sugieren un esquema de codificación donde se haga explícita la nomenclatura CMO en la lectura de transcripciones. El siguiente es un extracto de una entrevista hecha con una enfermera. Ambas oraciones fueron pronunciadas por el actor entrevistado. Se pone primero el ejemplo sin ninguna codificación para ilustrar el proceso.

> Ellos [los padres] saben cómo comunicarse, cómo actuar si el niño está llorando. Les estamos dando herramientas y se usan estas herramientas, que no tenían cuando llegaron al programa.
>
> Muchas de nuestras madres son madres primerizas, tenemos gente que vienen por segunda vez, pero creo que se beneficia de tener allí a los padres comunitarios. Es que la persona con quien es más cómodo hablar [...] hablan su idioma o vienen del país de donde usted viene (Jackson y Kolla, 2012, p. 343).

En las siguientes etapas del análisis (acá solo se presenta un resumen), las investigadoras agregan, por tanto, las nomenclaturas correspondientes a Contextos, Mecanismos y Resultados, utilizando los siguientes criterios: C se utiliza para caracterizar algo (ej. del lugar o de las personas) existente antes del programa; M busca recoger actividades o acciones emprendidas por actores del programa y O captura resultados obtenidos a partir de la interacción de acciones o actividades en situaciones particulares. Un posible resultado de dicho ejercicio sería el siguiente:

> Ellos [los padres] saben cómo comunicarse, cómo actuar si el niño está llorando **[O: buenas interacciones entre padres e hijos]**. Les estamos dando herramientas y se usan estas herramientas **[M: enseñanza de herramientas y capacidades para fortalecer vínculos]**, que no tenían cuando llegaron al programa **[C: El conocimiento de los padres sobre herramientas para interactuar con sus hijos]**
>
> Muchas de nuestras madres son madres primerizas **[C: participantes son padres por primera vez]**, tenemos gente que vienen por segunda vez **[O: uso frecuente del programa]**, pero creo que se beneficia de tener allí a los padres comu-

> nitarios **[O: cumplimiento objetivo específico del programa]**. Es que la persona con quien es más cómodo hablar **[M: es cómodo, confiable]** ... hablan su idioma o vienen del país de donde usted viene **[M: hay entendimiento cultural]** (Jackson y Kolla, 2012, p. 343).

Este segundo ejemplo puede ser aprovechado para hacer una advertencia sobre la lógica general del proceso de extracción de resultados. La razón para citarlo es que sin duda ilustra una lógica operativa concreta sobre cómo leer una pieza de información cualitativa –algo que muchos evaluadores le reclaman a quienes producen textos metodológicos– desde una óptica realista. En tal sentido, y desde una perspectiva netamente operativa, repetir este ejercicio con algunas entrevistas clave puede servir como un mecanismo heurístico útil para ejercitar la reflexividad del investigar en torno a la lectura del material recogido. Sin embargo, una ER rigurosa requiere evitar caer en lo que Pawson y Manzano (2012) denominan un realismo cualitativo, o la práctica de seleccionar pedazos de evidencia que resalten narrativas individuales sobre experiencias positivas o negativas de una intervención sin el rigor suficiente que indique que se ha identificado un patrón en los datos. Este es un mensaje fundamental, y por tanto vale la pena ser reiterativos al respecto:

> La investigación realista explica patrones y estos no pueden ser determinados a partir de apuntes anecdóticos (por parte de sujetos) o de pensar con el deseo (por parte de evaluadores). Los resultados deben ser cuidadosamente conceptualizados y sus indicadores bien pensados; deben establecerse líneas de base; mediciones del antes y el después deben ser trazadas; cohortes completas de sujetos deben ser monitoreadas. Ninguno de estos requisitos requiere una reorientación fundamental de la estrategia de investigación. Muchos de ellos son también los prerequisitos de buenos datos administrativos [...]. El punto clave es referirse a la teoría [del programa] (Pawson y Manzano, 2012, p. 183).

En sí, el mensaje es el mismo: todo depende de las configuraciones CMO y la búsqueda de refinarla tanto como sea posible —según tiempo, presupuesto, logística—. Ello quiere decir que los ejemplos acá presentados no pueden (ni deben) ser tomados como una única receta para hacer una buena evaluación realista. Sin embargo, se ponen sobre la mesa en tanto son muestras palpables de que es posible poner en práctica esquemas de pensamiento complejo y de triangulación de información –desde la recolección misma de datos, y no solo, aunque también,

sobre datos consolidados– al servicio de la evaluación de intervenciones en todo tipo de sectores económicos, políticos y sociales.

V. COMENTARIO FINAL: NO HAY QUE TENER MIEDO A EMPEZAR

En este artículo se ha puesto la ambiciosa tarea de introducir un esquema de evaluación que permita avanzar hacia el cumplimiento de algunas de las promesas inconclusas del paradigma de evaluación dominante. Las técnicas de evaluación de impacto –que suelen ser presentadas como el aparato metodológico de la política basada en la evidencia– son útiles para tareas de medición de efectos, pero, como lo reconocen de manera explícita algunos de sus principales exponentes, cuentan con limitaciones concretas al momento de ofrecer explicaciones del por qué y el cómo funciona o no una intervención o programa y bajo qué circunstancias es más factible que muestre resultados. Si dicho diagnóstico es cierto, surge necesariamente la pregunta sobre el alcance o la utilidad de las técnicas experimentales de evaluación para informar la tarea del hacedor de política.

Al discutir lo anterior algunos lectores podrían sentir que las ideas expuestas en este documento tienen como objetivo deslegitimar el uso de herramientas experimentales. Dado que existe el riesgo que se caiga en dicha simplificación, vale la pena hacer algunas aclaraciones adicionales. La primera —y más relevante quizás– es que en ningún momento se ha propuesto abandonar la cuantificación, en tanto se reconoce como una práctica con gran valor metodológico en la actividad evaluadora. En el párrafo anterior ya se hizo mención de esta idea –en particular, y en sintonía con argumentos esbozados en este artículo, porque la estadística probabilística juega un papel crucial al momento de orientar al evaluador sobre posibles tendencias y regularidades que indican la operación de mecanismos sociales–. De hecho, aunque no fue foco de la discusión entablada en el texto, el realismo como postura filosófica se aleja de argumentos que tienden a desechar lo cuantitativo (sencillamente) porque es arbitrario o porque no existe tal cosa como el ser objetivo. Los números importan, pero el punto es que su condición ontológica –empirista, atomista– hace que sean insuficientes para explicar la emergencia y la transformación de fenómenos sociales.

Una segunda aclaración importante para evitar falsos antagonismos es que la ER y las evaluaciones experimentales, desde sus bases conceptuales y epistemológicas,

comparten la motivación de entender y de explicar. Este no es el caso de otras corrientes de pensamiento, como las tendencias posmodernistas o constructivistas (extremas). De hecho, la consigna de estas últimas es mucho más de tipo relativista, paradigma dentro del cual no cabe la noción de verdades objetivas y conexiones causales entre fenómenos sociales (Bhaskar, 1998; Sayer, 2000; Parra, 2017). El punto por resaltar es que el realismo tiene fines tácitos más cercanos a los experimentadores que al de sus críticos, lo cual hace su crítica (potencialmente) más plausible para realmente mejorar la capacidad de las herramientas de evaluación en su tarea de proveer recomendaciones que ayuden a transformar entornos sociales.

Pese a lo anterior, algunos lectores podrían insistir en que los métodos de evaluación experimental ofrecen resultados más prácticos y, por tanto, sujetos a implementación. En tal caso, sería interesante consultar la obra de Mario Bunge (2011), filósofo y físico, quién una vez sostuvo que la consistencia, la sofisticación y la belleza formal nunca son suficientes en la investigación científica, cuyo producto final se espera que coincida con la realidad. El mundo social es complejo y, por tanto, las herramientas que deben utilizarse para estudiarlo deben ser capaces de capturar y de lidiar, tanto como sea posible, con dicha complejidad. Simplificar es importante, en tanto el no hacerlo podría conducir a la trampa de aquel emperador chino en la obra de Borges que quiso construir un mapa del imperio con tal grado de detalle que ya no le fue útil para tomar ninguna decisión. Sin embargo, parafraseando a Einstein, una explicación debe hacerse tan simple como sea posible, pero no más simple que ello. Es decir, y en línea con los postulados realistas, el arte de simplificación no puede omitir lógicas básicas sobre la forma en que individuos, grupos y sociedades se relacionan (de forma no atomística) e interpretan su entorno y su propia historia.

Ante lo anterior, se debe evitar caer en interpretaciones imprecisas sobre las herramientas de rigor de la ER: no todo vale, en tanto ello implicaría contrariar postulados básicos del realismo en torno a la existencia y la posibilidad de estudiar objetos sociales (ej. programas) con poderes causales (Parra, 2016). Por motivos de enfoque y espacio, el presente artículo no ha discutido en detalle criterios de calidad de una evaluación realista. También se ha omitido, de momento, la presentación de avances de esta tradición para responder preguntas de interés para un planeador público, como por ejemplo el análisis costo-beneficio de una intervención. Los trabajos como el de Wong, *et al.* (2016) o el de Anderson y Hardwick (2016), o el sitio de web del proyecto Realist And Meta-narrative Evidence

Syntheses: Evolving Standards (RAMESES) son, por tanto, referencias obligatorias para profundizar en estas discusiones.

Para finalizar, vale la pena volver a citar a los precursores de la evaluación realista, esta vez para enfatizar en dos consejos prácticos para practicantes. El primero dice: no tenga miedo de hacer grandes preguntas sobre pequeñas intervenciones y de usar pequeñas intervenciones para probar grandes teorías (Pawson y Tilley, 2001, p. 322). El segundo, por su parte, sugiere utilizar múltiples métodos y múltiples fuentes de datos según la oportunidad y necesidad (Pawson y Tilley, 2001, p. 323). Estos son mensajes claros sobre la importancia del verdadero trabajo interdisciplinario al servicio del debate sobre las políticas públicas y, por tanto, los rumbos de una sociedad. En la medida que el mundo es complejo, pero que también sea posible (como lo reitera el realismo) estudiarlo, no hay otra salida que el diálogo y el constante contraste entre visiones y teorías para descubrir y transformar mecanismos sociales que afectan la vida de individuos y colectividades. En este sentido, y contrario a lo que lo que algunos críticos podrían suponer *a priori*, la posición a favor de la ciencia que evoca el realismo es también un llamado a la inclusión y, si se quiere, a la democracia.

REFERENCIAS

Anderson, Rob, and Rebecca Hardwick (2016), "Realism and Resources: Towards more Explanatory Economic Evaluation", *Evaluation*, Vol. 22, No. 3.

Bhaskar, Roy (1998), *The Possibility of Naturalism. A Philosophical Critique of the Contemporary Human Sciences*, London and New York: Routledge.

Blamey, Avril, y Mhairi Mackenzie (2007), "Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges?", *Evaluation*, Vol. 13, No. 4.

Bozzoli, Carlos, Tilman Brück, and Nina Wald (2013), "Evaluating Programmes in Conflict-Affected Areas", en Patricia Justino, Tilman Brück, and Philip Verwimp (editors), *A Micro-Level Perspective on the Dynamics of Conflict, Violence, and Development*, Oxford: Oxford University Press.

Bunge, Mario (2011), "Knowledge: Genuine and Bogus", *Science & Education*, Vol. 20, No. 5.

Bush, Kenneth, and Collen Duggan (2013), "Evaluation in Conflict Zones: Methodological and Ethical Challenges", *Journal of Peacebuilding & Development*, Vol. 8, No. 2.

Deaton, Angus (2010), "Instruments, Randomization, and Learning about Development", *Journal of Economic Literature*, Vol. 48, No. 2.

Emmel, Nick (2013), *Sampling and Choosing Cases in Qualitative Research*, Los Angeles, London, New Delhi, Singapur, Washington: SAGE.

Gertler, Paul, Sebastian Martinez, Patrick Premand, Laura B. Rawlings, y Christel M. J. Vermeersch (2011), *Evaluación de impacto en la práctica*, Washington: The World Bank.

Glewwe, Paul, Eric Hanushek, Sarah Humpage, and Renato Ravina (2013), "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010", in Paul Glewwe (editor), *Education Policy in Developing Countries*, Chicago: University of Chicago Press.

Guba, Egon, and Yvonna S. Lincoln (1989), *Fourth Generation Evaluation*, Newbury Park, London, New Delhi: SAGE Publications.

Hesse-Biber, Sharlene (2013), "Thinking Outside the Randomized Controlled Trials Experimental Box: Strategies for Enhancing Credibility and Social Justice", *New Directions for Evaluation*, No. 138.

Horrocks, Ivan, and Leslie Budd (2015), "Into the Void: A Realist Evaluation of the eGovernment for You (EGOV4U) Project", *Evaluation*, Vol. 2, No. 1.

Jackson, Suzanne, and Gillian Kolla (2012), "A New Realistic Evaluation", *American Journal of Evaluation*, Vol. 33, No. 3.

Manzano, Ana (2011), "A Realistic Evaluation of Fines for Hospital Discharges: Incorporating the History of Programme Evaluations in the Analysis", *Evaluation*, Vol. 17, No. 1.

Manzano, Ana (2016), "The Craft of Interviewing in Realist Evaluation", *Evaluation*, Vol. 22, No. 3.

Milani, Carlos (2009), "Evidence-Based Policy Research: Critical Review of Some International Programmes on Relationships between Social Science Research and Policy-Making", *Policy Papers*, No. 18, United Nations Educational Scientific and Cultural Organisation (UNESCO).

Montoya-Vargas, Juny (2014). "The Field of Curriculum Studies in Colombia", in William F. Pinar, (editor), *International Handbook of Curriculum Research*, Vol. 2, New York: Routledge.

Parada, Jairo (2007), "Sociedad y evaluación de programas sociales en el realismo crítico: una revisión crítica", *Investigación y Desarrollo*, Vol. 15, No. 1.

Parra, Juan David (2013), "Preferencias endógenas, prosocialidad y políticas públicas", *Divergencia*, No. 15.

Parra, Juan David (2015), "The Paradigm of Critical Realism and Involving Educators in Policy Debates", *Gist: Education and Learning Research Journal*, No. 10.

Parra, Juan David (2016), "Realismo crítico: Una alternativa en el análisis social", *Sociedad y Economía*, No. 31.

Parra, Juan David (2017), "Critical Realism and School Effectiveness Research in Colombia: The Difference it Should Make", *The British Journal of Sociology of Education*, en imprenta.

Patomäki, Heikki (2003), "A Critical Realist Approach to Global Political Economy", in Justin Cruickshank (editor), *Critical Realism: The Difference It Makes*, London: Routledge.

Pawson, Ray (1996), "Theorizing the Interview", *The British Journal of Sociology*, Vol. 47, No. 2.

Pawson, Ray (2003), "Nothing as Practical as a Good Theory", *Evaluation*, Vol. 9, No. 4.

Pawson, Ray (2006), *Evidence-Based Policy: A Realist Perspective*, London: Sage.

Pawson, Ray (2013), *The Science of Evaluation: A Realist Manifesto*, London: Sage.

Pawson, Ray, and Ana Manzano (2012), "A Realist Diagnostic Workshop", *Evaluation*, Vol. 18, No. 2.

Pawson, Ray, and Nick Tilley (1997), *Realistic Evaluation*, London, Thousand Oaks, New Delhi: SAGE Publication.

Pawson, Ray, and Nick Tilley (2001), "Realistic Evaluation Bloodlines", *American Journal of Evaluation*, Vol. 22, No. 3.

Pawson, Ray, Geoff Wong, and Lesley Owen (2011a), "Known Knowns, Known Unknowns, Unknown Unknowns: The Predicament of Evidence-Based Policy", *American Journal of Evaluation*, Vol. 32, No. 4.

Pawson, Ray, Geoff Wong, and Lesley Owen (2011b), "Myths, Facts and Conditional Truths: What is the Evidence on the Risks Associated with Smoking in Cars Carrying Children?", CMAJ, Vol. 183, No. 10.

Porpora, Douglas (2015), *Reconstructing Sociology. The Critical Realist Approach*, Cambridge: Cambridge University Press.

Porter, Sam (2015), "The Uncritical Realism of Realist Evaluation", *Evaluation*, Vol. 21, No. 1.

Sayer, Andrew (2000), *Realism and Social Science*, Londres, Thousand Oaks and New Delhi: SAGE.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalised Causal Inference*, Boston and New York: Houghton Mifflin.

Smith, Chris, and Tony Elger (2014), "Critical Realism and Interviewing Subjects", in Paul Edwards, Joe O'Mahoney, and Steve Vincent (editors), *Studying Organizations Using Critical Realism. A Practical Guide*, Oxford: Oxford University Press.

Sridharan, Sanjeev, and April Nakaima (2011), "Ten Steps to Making Evaluation Matter", *Evaluation and Program Planning*, Vol. 34, No. 2.

Stake, Robert E. (2004), *Standards-Based & Responsive Evaluation*, Thousand Oaks, London and New Delhi: SAGE.

Starr, Martha (2014), "Qualitative and Mixed-Methods Research in Economics: Surprising Growth, Promising Future", *Journal of Economic Surveys*, Vol. 28, No. 2.

van Belle, Sara, Geoff Wong, Gill Westhorp, Mark Pearson, Nick Emmel, Ana Manzano, and Bruno Marchal (2016), "Can 'realist' Randomised Controlled Trials be Genuinely Realist?", *Trials*, Vol. 17, No. 1.

Vygotsky, Levy (1997), "Interaction between Learning and Development", in Mary Gauvain, and Michael Cole(editors), *Readings on the development of children*, New York: W.H. Freeman and Company.

Westhorp, Gill (2013), "Developing Complexity-Consistent Theory in a Realist Evaluation", *Evaluation*, Vol. 19, No. 4.

Wong, Geoff, and Chrysanthi Papoutsi (2016), *Notas de clase (14-18 de Octubre): Realist Reviews and Realist Evaluation*, Oxford: Oxford University

Wong, Geoff, Gill Westhorp, Ana Manzano, Joanne Greenhalgh, Justin Jagosh, and Trish Greenhalgh (2016), "RAMESES II Reporting Standards for Realist Evaluations", *BMC Medicine*, Vol. 14, No. 96.