UNED Facultad de Ciencias Económicas y Empresariales Departamento de Economía Aplicada y Estadística

Revista de Evaluación de Programas y Políticas Públicas
JOURNAL OF PUBLIC PROGRAMS AND POLICY EVALUATION

EVALUACIONES DE IMPACTO EN LA COOPERACIÓN PARA EL DESARROLLO

IMPACT EVALUATIONS IN INTERNATIONAL DEVELOPMENT CO-OPERATION

Núm. 3 (2014), pp. 117-153.

José María Larrú [1].

Recibido: **Junio, 2014**

Aceptado: **Septiembre, 2014**

JEL Clasif: C93; F35

[1]. Universidad CEU San Pablo, larram@ceu.es

Resumen

Los diseños experimentales están actualmente en la frontera metodológica de la evaluación de impacto y de la economía del desarrollo. Se enfrentan a críticas, algunas generales a toda evaluación, otras específicas de su diseño aleatorizado. El trabajo señala y discute cuatro de ellas: la necesidad de teoría para diseñar creativamente cada evaluación; el problema de la heterogeneidad; la validez externa; y la traslación de la evidencia a las políticas públicas. A la luz de la experiencia española con el Fondo Español de Evaluaciones de Impacto, la principal recomendación que surge es que las Administraciones Públicas, deben tener conocimiento de las potencialidades y limitaciones de la amplia gama de evaluaciones de impacto para apoyar su ejecución cuando sea oportuno y para ajustar la pregunta evaluativa que encarguen a las condiciones de rigor, tiempo, presupuesto y utilidad de la información que recibirán de cara a tomar una decisión política.

Palabras Clave: cooperación al desarrollo; evaluación; experimentos aleatorios controlados; validez externa.

Abstract

Experimental designs are currently in the methodological frontier of impact assessment and development economics. The methodology has received some criticisms: some of them are applied to any type of evaluation while others are specific to their randomized design. The paper points out and discusses four of them: the need for theory to design creatively each evaluation; the problem of heterogeneity; external validity; and the translation of evidence into public policy. In light of the Spanish experience with the Spanish Impact Evaluation Fund, the main recommendation that emerges from the analysis is that policy makers should be aware of the potential and limitations of the wide range of impact assessments, to support their implementation when appropriate and to adjust the inquiry to the necessary accuracy, time, budget constraints and to the utility of the information they receive in order to take a political decision.

Key Words: international development co-operation; evaluation; random control trials; external validity.

1. Introducción.

En una revisión de 80 evaluaciones de cooperación para el desarrollo publicadas en DEReC[2] entre 2004 y 2007 y promovidas por 22 donantes, Forss & Bandstein (2008) encontraron las siguientes características significativas:

- un tercio eran evaluaciones de proyecto y dos tercios de programas, a pesar del énfasis puesto en los discursos sobre nuevos instrumentos (apoyo presupuestario, sectorial, etc.);
- el 10% no eran realmente evaluaciones sino informes anuales, documentos de política, resúmenes, auditorias de resultados, entre otros;
- dos tercios habían sido realizadas mientras la intervención seguía llevándose a cabo;
- el 66% evaluó eficacia y pertinencia y el 50% impacto, eficiencia y sostenibilidad;
- los diseños eran todos muy semejantes y sólo en un caso (de 80) fue experimental y en el 5,2% fue cuasi-experimental;
- en sólo un caso se trató el problema del sesgo de selección y estimación de spillovers y en menos del 7% de los casos hubo tratamiento de contrafactuales;
- en el 90% de los casos se emplearon análisis documental y entrevistas, en el 50% observación y visita al terreno como fuentes de información y en el 15% cuestionarios (casi siempre coincidiendo con los diseños cuasi-experimentales);
- casi sin excepción los informes no detallan el tipo de muestreo empleado, anexan el cuestionario concreto o justifican la selección de los entrevistados.

Para buscar la causa de este tipo de caracterización, los autores investigan los términos de referencia (TdR) de las evaluaciones seleccionadas. Esta revisión les conduce a la conclusión de que la mayoría de los TdR solicitan algún tipo de "resultados", pero que éste término es polisémico, pues en algunos casos se está

[2] DAC Evaluation Resource Centre: http://www.oecd.org/derec/

solicitando productos (*outputs*), en otros efectos (*outcomes*), en otros impactos (*impacts*) y en varios de ellos, alguna combinación imprecisa de los tres. Una primera conclusión interesante del estudio es que *resultados* debe ser un término "contexto-específico" en el que se detalle en cada caso qué se entiende por una evaluación de resultados.

Un diagnóstico similar podría hacerse del caso español. Si, por una parte, el Informe Anual de la Evaluación en la Cooperación Española 2010 (MAEC 2012) identifica mediante encuesta a actores más de 500 evaluaciones en la Cooperación Española, ninguna podría ser calificada de evaluación de impacto dentro de lo que el ámbito académico entiende por tal[3]. Es posible que dentro de esas evaluaciones se encuentren juicios sobre los impactos, ya que este es uno de los criterios "clásicos" o canónicos del CAD junto a la pertinencia, eficacia, eficiencia y sostenibilidad, pero que no se corresponde con el concepto "académico" de evaluación de impacto que lo liga indefectiblemente a concretar el cambio producido en una población si no se hubiera intervenido en ella (lo que se conoce como contrafactual). Por otra parte, el vigente documento sobre la "Política de Evaluación de la Cooperación Española" (MAEC 2013a) reconoce el deseo de ir pasando "de las evaluaciones centradas en las actividades y productos, a las de resultados e impactos en el medio y largo plazo" (MAEC 2013:IX punto 13) dentro de la gestión orientada a resultados[4]. Sin embargo, apenas una de las más de 400 evaluaciones seleccionadas en el Plan Bienal de Evaluaciones 2013-2014 (MAEC 2013b) es calificada como evaluación de impacto[5].

Otro rasgo que justifica la caracterización realizada arriba, que bien podría generalizarse como la tendencia actual en las evaluaciones de cooperación para el

[3] Al no estar publicadas la inmensa mayoría de ellas, no es posible hacer un diagnóstico más preciso sobre esta caracterización. En la actualidad, son fácilmente accesibles on line (www.cooperacionespanola.es) los 38 informes de evaluación promovidos por la División de Evaluaciones de la Secretaría General de Cooperación Internacional para el Desarrollo (SGCID) del Ministerio de Asuntos Exteriores y Cooperación. Ninguno de ellos es una evaluación de impacto.

[4] "En el ámbito de la cooperación para el desarrollo, la gestión orientada a resultados y la búsqueda de una mayor rendición de cuentas están desplazando el foco de atención de manera creciente hacia la evaluación de los resultados e impactos" (MAEC 2013:5, punto 32).

[5] Cabe destacar la realización de una evaluación del impacto del Programa Cisternas llevado a cabo en Brasil en el marco del FCAS" [Fondo de Cooperación para Agua y Saneamiento], (MAEC 2013b:5). Tengo constancia de la intención por parte de la ONGD "Alianza por la Solidaridad" dentro un convenio financiado por AECID en tres países de África Subsahariana para la implantación de cocinas mejoradas, de llevar a cabo una evaluación que sí podría considerarse como una verdadera evaluación de impacto. Si esta iniciativa logra realmente llevarse a cabo, podría ser la primera evaluación de impacto "pura" de la Cooperación Española.

desarrollo, es que se piden demasiadas cosas a la vez, sin priorizar criterios (los cinco clásicos pueden pensarse como dimensiones de los resultados). Esta "ambición" condiciona mucho los diseños de las evaluaciones, restringiendo el campo de los diseños cuasi-experimentales y –en el margen- de los experimentales aleatorios, dada su necesidad de que estén concebidos antes de que la intervención comience y que responden a eficacia e impacto, pero no a pertinencia, eficiencia o sostenibilidad.

Forss & Bandstein (2008) mencionan también la multi-intencionalidad de los informes. En 15 evaluaciones se menciona la comparación como fin primario, en 28 generar soporte para la toma de decisiones y en 19 el aprendizaje. Por el contrario, en 10 evaluaciones no se fue capaz de detectar su finalidad, más allá de recabar información, pero sin saber cómo y para qué será usada. Esta es otra característica de la Cooperación Española. Muy a menudo, las evaluaciones no son lideradas por el responsable político de la toma de decisiones. No hay una involucración real, con una pregunta concreta (o hipótesis) que desee ser respondida para tomar la decisión mejor, basándose en la información aportada por la evaluación (Larrú 2011, 2012, Larrú y Méndez 2012). Más bien el perfil suele ser evaluaciones de "rendición de cuentas", de revisión sobre lo que se ha hecho bajo algún programa o en un país, pero no se definen como "de impacto" para conocer con precisión qué estrategia es la más eficiente entre varias alternativas, o cuál es el "efecto neto" atribuible y causado por la intervención.

En resumen, hasta hace muy poco, la inmensa mayoría de las evaluaciones promovidas en el ámbito de la cooperación al desarrollo han estado orientadas a la gestión y no al conocimiento preciso de si las intervenciones funcionan o no, ni por qué lo hacen o cuál es su coste comparativo entre alternativas (eficiencia). Existe un predominio de preguntas sobre la gestión. Por ejemplo, cómo se utilizaron los fondos asignados, grado de cumplimiento de las actividades y productos [*outputs* no *outcomes*] programadas, sistematización de lugares y programas que han sido financiados a lo largo de una región y tiempo, con cierto juicio sobre su eficacia o eficiencia pero sin pretender un análisis de atribución de los cambios observados. Además, habría que añadir ofertas de tiempo muy cortas (nunca más de un año). Este panorama hace que algunos analistas españoles se muestren pesimistas sobre el pasado, presente y futuro de las evaluaciones de impacto y los diseños experimentales, a pesar de la "moda" internacional de los experimentos aleatorios controlados (*Random Control Trials*, RCTs por sus siglas inglesas).

Sin embargo, España pudo haber sido una de las excepciones en el panorama hasta aquí descrito. En 2007 promovió la creación del Fondo Español para Evaluaciones de Impacto (*Spanish Trust Fund for Impact Evaluation, SIEF)* ubicado dentro de la Red de Desarrollo Humano del Banco Mundial. España comprometió 10,4 millones de Euros en este Fondo hasta 2010[6]. En 2012, fruto –entre otras razones, como veremos más adelante- de los recortes presupuestarios en la cooperación española al desarrollo, no se pudo continuar con la financiación. Actualmente, una segunda versión del SIEF –ahora denominado *Strategic Impact Evaluation Fund* y sostenido financieramente por la cooperación británica, ha tomado el relevo hasta 2018[7].

Pero las evaluaciones de impacto, y más concretamente las que se realizan bajo los diseños experimentales, en particular los RCTs, no están exentos de debate académico, ni político. Frente a grupos impulsores y acérrimos defensores de sus ventajas (el Poverty Action Lab del MIT o "Ideas for Development" liderado por Karlan), están apareciendo buenos trabajos académicos (Blattman 2008; Rodrik 2008; Scriven 2008; Deaton 2010; Camfield & Duvendack 2014) que señalan puntos débiles, junto a declaraciones institucionales como la de la Asociación Europea de Evaluación (European Evaluation Society 2007) que, sin menospreciar el lugar que deben tener los RCTs, no les dan ninguna primacía sobre el resto de metodologías evaluativas.

En este trabajo se señalan algunas de estas deficiencias o limitaciones que tienen los diseños experimentales y cuál está siendo la respuesta de sus defensores (por ejemplo Banerjee 2006; Banerjee & Duflo 2008; Easterly 2008; White 2014).

En concreto se analiza más en profundidad la necesidad de teoría para diseñar creativamente cada evaluación (no se deben "rutinizar" mecánicamente), el problema de la heterogeneidad (se mide el efecto promedio entre los tratados –intencionada o

[6] Un subproducto l –casi el único beneficio obtenido por la Cooperación Española- fue la celebración de un curso de formación de cinco días (junio de 2008) en Madrid, sobre cómo llevar a cabo este tipo de evaluaciones, cómo promoverlas y difundirlas. En abril de 2012 tuvo lugar una Jornada sobre evaluación de impacto que sirvió de cierre a esta iniciativa. Para mayor detalle sobre el funcionamiento del SIEF véase Larrú (2010) y la evaluación externa llevada a cabo por Feinstein (2012), disponible en el sitio web: http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/ORGANIZATION/EXTHDNETWORK/EXTHDOFFICE/0,,contentMDK:22390002~menuPK:6575735~pagePK:64168445~piPK:64168309~theSitePK:5485727,00.html .

[7] Este SIEF 2.0 está llevando a cabo evaluaciones en educación, nutrición infantil, salud, agua y saneamiento en 24 países. Para los detalles de este Fondo 2012-2018, puede consultarse su sitio web: http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/ORGANIZATION/EXTHDNETWORK/EXTHDOFFICE/0,,contentMDK:23147035~menuPK:6508083~pagePK:64168445~piPK:64168309~theSitePK:5485727,00.html

realmente- pero no el efecto individual o sobre subgrupos de participantes, aunque esto último podría hacerse si se tiene en cuenta en el diseño inicial), la validez externa (los problemas de "escalabilidad" o replicación en diferentes contextos de espacio y tiempo no tiene por qué arrojar resultados idénticos) y la traslación de la evidencia a las políticas públicas para que realmente este tipo de evaluaciones, -rigurosas sí, pero también largas y costosas- puedan ser generadoras de programas de mejoras sociales concretos.

La complejidad contextual (político-social) del desarrollo humano (y económico) fuerza a tomar en cuenta otros métodos de evaluación, sobre todo cuando la información requerida se centra en cuestiones de gestión o de poder, pero que cuando la pregunta central es si una intervención de cooperación para el desarrollo (por extensión, una intervención social) funciona o no, es decir, cuando conocer la eficacia y el impacto es lo más importante, la primera opción metodológica a considerar debe ser el diseño experimental, sabiendo que no existe un "estándar dorado" (Deaton 2010) que deba menospreciar otras formas de conocer y evaluar. El papel de la teoría para responder a *por qué* funciona o no una intervención es esencial (White 2009) y no puede ser respondido por la técnica econométrica más depurada que podamos desarrollar o aplicar. En última instancia, una técnica econométrica altamente refinada, pero sin teoría económico-social o sin antropología y sin psicología social, puede aportar "un dato" o "el número", pero sin una explicación o historia convincente que lo interprete, ni un contexto que "moldee el modelo". El número (la diferencia de los promedios entre los grupos de tratamiento y control) no servirá para transformar la realidad de la pobreza y el subdesarrollo, que a menudo es más un problema político y social que técnico o incluso económico (mucho menos monetario, en el sentido de que encuentre solución sólo dando más dinero en forma de ayuda oficial).

El resto del trabajo se organiza de la siguiente manera. En la siguiente sección se hace un breve recorrido histórico sobre la expansión y "puesta de moda" de la evaluación de impacto, presentando los principales agentes académicos y grupos de evaluación que están centrados en la evaluación de impacto. En la tercera sección se abordan principales problemas metodológicos y epistemológicos que afectan a las evaluaciones de impacto. En la cuarta sección se resumen las principales conclusiones y aplicaciones de lo expuesto.

2. Antecedentes y agentes en la evaluación de impacto en cooperación al desarrollo.

A finales de los años '90 la respuesta más consensuada que se recibía al preguntar a los especialistas por el criterio de impacto incluido en el quinteto "canónico" del CAD era algo parecido a: "realmente a esto [del impacto] nadie le ha hincado el diente". Se asumía entonces el concepto de impacto como un aspecto evaluativo deseable (determinar los cambios de largo plazo, intencionados o no, directos e indirectos en la zona de intervención), pero casi imposible.

Este escenario comenzó a cambiar cuando investigadores como Abhijit Banerjee, Esther Duflo y Sendhil Mullainathan fundaron en 2003 el Poverty Action Lab, con el apoyo del departamento de Economía del MIT. Su empeño por superar el recelo de que no se podría "experimentar" con programas sociales, les llevó a conducir evaluaciones de impacto que –sin menospreciar nunca los aspectos éticos- les permitían realizar asignaciones aleatorias del beneficio de una intervención entre los elegibles.

En 2004, se habían incorporado al Laboratorio dos excelentes investigadores: Rachel Glennerster, procedente del FMI y con amplia experiencia en gestión, y Dean Karlan, que posteriormente será profesor en Yale y fundador del Innovations for Poverty Action (IPA), otro de los núcleos más activos de la actual corriente en torno a la evaluación de impacto[8]. El fin último del J-PAL queda claro en su lema: *translating research into action*. Probablemente el sueño de muchísimos (si no de todos) los investigadores del mundo.

La excelencia académica de Duflo (ganadora entre muchos otros, de la prestigiosa John Bates Clark Medal a la investigadora en economía más influyente menor de 40 años[9]) y probablemente la capacidad de traducir evaluaciones en artículos académicos de primer nivel como el de Miguel y Kremer (2004) sobre desparasitación en Kenia, junto a los propios de Duflo y Banerjee hicieron que el panorama en torno a la

[8] Una síntesis de la evolución histórica e impresionante desarrollo del Poverty Action Lab puede consultarse en su sitio web: http://www.povertyactionlab.org/es/historia. En la actualidad cuenta con 76 profesores afiliados, 373 evaluaciones en curso o completadas en 52 países y 1.512 personas capacitadas. En 2008 recibió el primer premio "Fronteras del Conocimiento" de la Fundación BBVA. El de 2013 ha recaído en uno de los socios indios más importantes en la construcción de evidencias en materia de educación por parte del J-PAL, la ONG India Pratham.

[9] Véase la descripción de su trayectoria en Udry (2011).

evaluación de impacto y a la economía del desarrollo cambiara y se convirtiera en una corriente quizá minoritaria en cuanto a la proporción de evaluaciones con diseño aleatorizado respecto al total de informes, pero muy activa y de influencia creciente. Intervenciones de Duflo en el Banco Mundial (Duflo 2004) o ante la *Econometric Society* (Duflo 2005), así como las primeras publicaciones que explicaban la posibilidad de aplicar diseños experimentales en las intervenciones de desarrollo (Duflo y Kremer 2005) fueron abonando el terreno para que, ya en 2006, se publicara el influyente informe "*When Will We Ever Learn? Improving Lives through Impact Evaluation*" que era el resultado de numerosas reuniones previas de académicos y políticos de alto nivel de países en desarrollo coordinadas por el Center for Global Development, un reconocido *think tank* de Washington.

El informe partía de la constatación de que entre las 127 evaluaciones revisadas por la Organización Internacional del Trabajo, sólo dos contenían un diseño cuyas conclusiones eran rigurosas en términos de impacto (ILO 2002). Una revisión sobre evaluaciones en el sector de la salud llevada a cabo por Levine (2004), logró documentar 56 informes, pero tuvo que eliminar 27 de ellos por su incapacidad para documentar impactos. La revisión realizada por Victoria (1995) sobre las evaluaciones de UNICEF en 1992-93 encontró que sólo 44 de 456 informes contenían juicios que incluyeran realmente el impacto[10]. Tras la diseminación del informe "Cuándo aprenderemos", se creó una iniciativa que es clave para entender la evolución y desarrollo de las evaluaciones de impacto: la *International Initiative for Impact Evaluation* (conocida como 3ie, http://www.3ieimpact.org/) y que ha sido presidida desde Nueva Dehli por Howard White, un prestigioso investigador doctorado en Oxford, profesor invitado del ISS holandés, profesor visitante del IDS de Brighton, forjado en la polémica empírica de la eficacia de la ayuda de finales de los '90 y por aquel entonces evaluador del Grupo de Evaluación Independiente del Banco Mundial[11]. Es muy probable que este recorrido por los mejores centros de análisis del desarrollo internacional, hayan influido fuertemente en que en la 3ie no se defienda un "fundamentalismo experimental" que considere exclusivamente los *Random Control Trials* como la única metodología válida para hacer evaluaciones de impacto. Será este

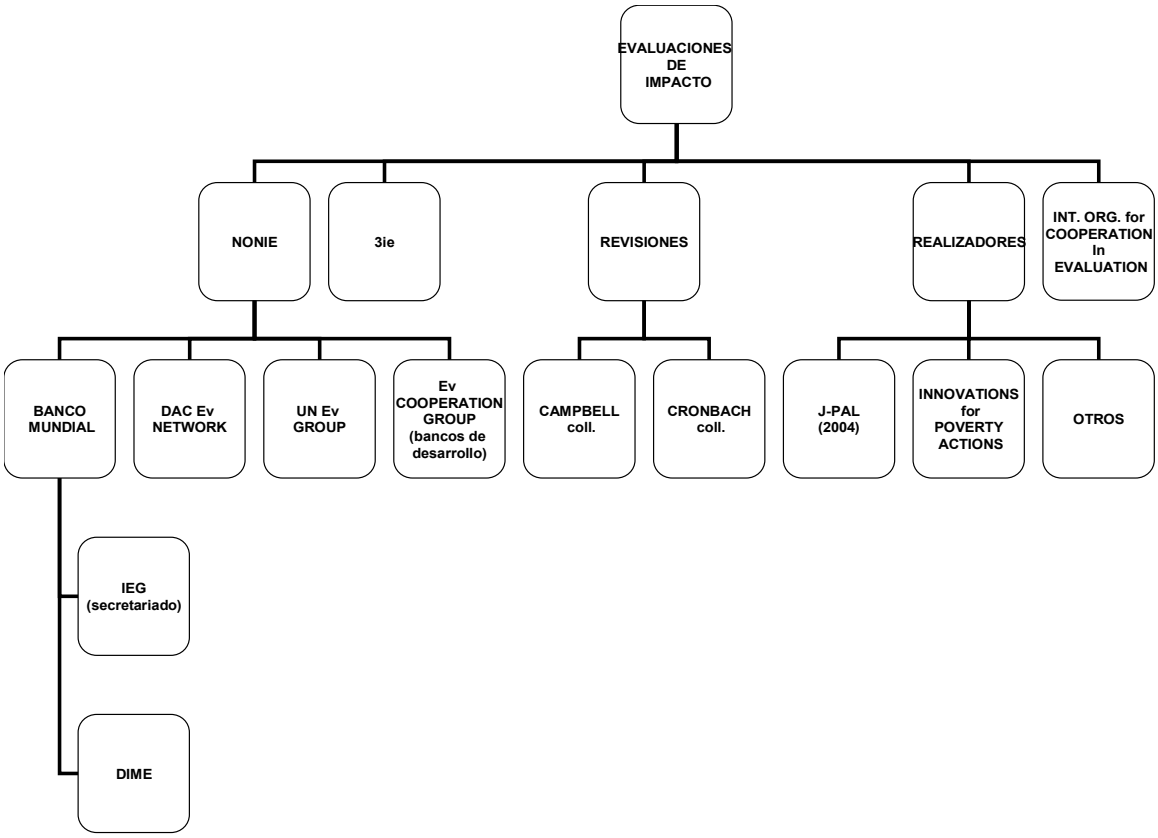
[10] Una de las actividades de la citada revisión fue la clasificación por parte de expertos de los informes de evaluación en posesión de UNICEF. El resultado fue que sólo el 20% de los informes considerados como evaluaciones de impacto lo eran realmente.

[11] Emmanuel Jiménez relevó en el cargo de director de la 3ie a Howard White en julio de 2014.

fundamentalismo adoptado por el Centro de Evaluación de Política Educativa Estadounidense, que anunció que sólo iba a financiar evaluaciones aleatorizdas, lo que provocó la reacción de la Asociación Europea de Evaluación a la que hemos hecho referencia en la introducción, así como de la homóloga Asociación Americana[12].

Hoy en día, son numerosas las instituciones que realizan, promueven o difunden evaluaciones de impacto o informes de síntesis sistemáticas a partir de ellas. Una visión sintética se ofrece en el Cuadro 1.

Cuadro 1. Organismos e instituciones relacionados con la evaluación de impacto.



[12] Véase American Evaluation Association (2003). Como muestra el documento, el fondo de la cuestión es la pretensión de ciertos evaluadores –para mí nunca aplicable a los autores reunidos en el J-PAL- de que los RCT sean la única forma de identificar la causalidad, el rigor y por tanto la única base para una toma de decisiones política válida, menospreciando métodos alternativos como la doble diferencia, el pareamiento, el uso de variables instrumentales o la discontinuidad en la regresión.

El primer organismo que muestra la figura, es la "red de redes", *Network of Networks for Impact Evaluation* (NONIE) creada en 2006 y compuesta por agencias bilaterales y multilaterales de cooperación al desarrollo. La NONIE se alimenta a su vez de cinco redes o fuentes de evaluaciones: i) el grupo de evaluaciones de cooperación internacional (http://www.ECGnet.org) compuesto por los bancos regionales de desarrollo (europeo, asiático, africano e interamericano e islámico), la oficina de evaluación independiente del FMI, la del Banco Europeo de Inversiones, la de la FAO y el IEG del Banco Mundial; ii) el grupo de evaluación de Naciones Unidas (http://www.uneval.org/); iii) la red de evaluaciones del Comité de Ayuda al Desarrollo de la OCDE, que corresponde a la cooperación bilateral, (http://www.oecd.org/dac/evaluation/); y iv) dos grupos del Banco Mundial: el de evaluaciones independientes (IEG) (http://ieg.worldbank.org/) y el de evaluaciones de impacto (DIME)[13].

En segundo lugar, la figura recoge la ya mencionada *International Initiative for Impact Evaluation* (http://www.3ieimpact.org/) surgida tras el impulso en torno al informe "*When Will Ever Learn*". En la actualidad esta Iniciativa lleva a cabo importantes contribuciones en el campo de unir las evaluaciones con las decisiones políticas, realizar informes de síntesis, convocar concursos y licitaciones abiertas de evaluaciones de impacto (ha financiado ya más de 100) y actualizar de forma permanente un repositorio de casi 2.400 informes de evaluación de impacto. Cuenta con un importante patrocinio de la agencia británica de ayuda y de la Fundación Bill y Melinda Gates.

Con un alcance más modesto en recursos, existen además algunas iniciativas regionales como el *Impact Evaluation Network* del LACEA latinoamericano (http://www.depeco.econo.unlp.edu.ar/cedlas/ien/about.htm).

Además de las redes y grupos multilaterales de evaluación de impacto, existen dos entidades especializadas en informes de síntesis sistemáticas a partir de informes evaluativos y que llevan los nombres de dos de los autores más señalados en el campo de la evaluación: la Cochrane (http://www.cochrane.org/) y la Campbell Collaboration (http://www.campbellcollaboration.org/) que también organizan congresos anuales.

13 http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,menuPK:3998281~pagePK:64168427~piPK:64168435~theSitePK:3998212,00.html

La primera está más especializada en el campo de la salud mientras que la segunda en desarrollo internacional, además de criminología y educación.

Puede mencionarse también la asociación *International Organization for Cooperation in Evaluation* (http://www.ioce.net/) como plataforma abierta a la colaboración y apoyo mutuo en el mundo de la evaluación, no específicamente en cooperación internacional al desarrollo, pero que destaca por su carácter voluntario y el estándar profesional de sus servicios. Junto con UNICEF e involucrando a numerosas organizaciones de la sociedad civil y de gobiernos, han logrado formar la *EvalPartners*, una Iniciativa de Asociación de Evaluación Internacional (*The International Evaluation Partnership Initiative*) que ha establecido 2015 como el año internacional de la evaluación (http://www.mymande.org/evalyear) .

Por lo que respecta a quién hace evaluaciones de impacto, ya hemos destacado en la introducción el SIEF, pero las referencias más conocidas son el Poverty Action Lab que –a fecha de junio de 2014- ofrece un buscador de 173 evaluaciones identificables por tema, región o desafío de política y el IPA (http://www.poverty-action.org/) que también dispone de un buscador temático de evaluaciones.

Una vez presentados los principales actores en la siguiente sección se analizan algunas de las características más controvertidas de la metodología experimental y se exponen algunos ejemplos de resultados que han ofrecido este tipo de evaluaciones.

3. Antecedentes y agentes en la evaluación de impacto en cooperación al desarrollo.

Prácticamente desde su renovada consideración y puesta en práctica en evaluaciones de economía del desarrollo, los diseños experimentales han recibido críticas (véanse por ejemplo Basu 2005; Blatman 2008; Deaton 2006 y 2009; Rao 2008; Rodrik 2008; Harrison 2014; Picciotto 2014) o resaltado sus limitaciones (Guijt & Roche 2014; Lensink 2014; White 2014). Picciotto 2014 considera que –siendo puristas- no son evaluaciones ya que juzgan aspectos clave como la pertinencia. Camfield & Duvendack (2014) resaltan la falta de validez externa, dudas sobre si el procedimiento es realmente "doble ciego" en la línea de Scriven (2008), los aspectos éticos, la atrición y los efectos externos o spillovers, así como aspectos prácticos no

menores como la extensa cooperación interinstitucional que requieren, sobre todo por parte de organizaciones locales que son las que deben recoger cuidadosamente los datos durante un largo periodo. El Cuadro 2 recoge las principales, agrupándolas en dos categorías: las limitaciones generales y los problemas técnicos internos. Algunas de las limitaciones generales no son específicas de los diseños experimentales, sino extrapolables a otros diseños o el hecho mismo de evaluar (como las cuestiones éticas y la dimensión política de la evaluación). Quizá lo que mejor expresa esta idea de crítica general es la observación de Rao (2008:130) de que lo esencial es el tipo de pregunta que está tras la evaluación: "al comenzar una investigación (o evaluación podríamos añadir) la pregunta que queremos contestar debe conducir al método que ha de utilizarse y no al revés".

Cuadro 2. Limitaciones de los diseños experimentales.

LIMITACIONES GENERALES
Imposibles de realizar cuando el alcance de la intervención es el universo poblacional
Consideraciones éticas (no se da tratamiento al grupo de control)
Dimensión política: se requiere una respuesta rápida, se desea garantía de sostenibilidad, problemas para aceptar que a una parte de la población (control) no se la va "ayudar" en ese momento y con ese tratamiento
Validez externa (ignorancia del efecto si se aplica a mayor escala; diferencias contextuales; entorno en continuo dinamismo que impide la no contaminación)
Problemas frente a diseños alternativos que puedan responder a la sostenibilidad, a los procesos, a las causas múltiples de intervenciones complejas
Coste y tiempo altos
Complejidad técnica elevada que no hace fácil transmitir el diseño y método de obtención de los resultados a los tomadores de decisiones

LIMITACIONES EN LA PROPIA ALEATORIZACIÓN
Debe basarse en una teoría previa y acertar con el diseño concreto
Cambio de conducta ante la certeza de ser observado (efecto Hawthorne –en el grupo de tratamiento- y John Henry –en el grupo de control-)
Heterogeneidad del efecto dentro del grupo de tratamiento desconocida
Hay que esperar al final de la intervención para conocer el efecto (que puede ser negativo)
Problemas de atrición y contaminación elevados (ej. en programas de infraestructuras) y de efectos de equilibrio general

A la hora de valorar un diseño concreto, debemos tener en cuenta también los costes de oportunidad. Las limitaciones generales aquí señaladas son igualmente aplicables a los diseños cuasi-experimentales (complejidad técnica, las dimensiones políticas, el coste y tiempo elevados son compartidos por el pareamiento o *matching*, a la discontinuidad en la regresión o la doble diferencia). Por eso, el resto de este trabajo no va a tratar estos problemas generales, sino que se centrará sólo en algunos de los problemas específicos que plantea la aleatorización, como es la heterogeneidad y la capacidad de atribuir causalidad en los diseños experimentales. Posteriormente se abordarán dos problemas comunes a toda evaluación: la necesidad de una teoría subyacente al programa o intervención y la economía política para avanzar en la escalabilidad y alcance del programa evaluado.

El problema de la heterogeneidad dentro del grupo de tratamiento.

Una de las mayores ventajas que tiene el diseño experimental es que elimina el sesgo de selección, sobre todo cuando hay auto-selección para participar en un programa social. Esto es básico en las intervenciones de cooperación al desarrollo, donde precisamente la elegibilidad se “sesga” hacia los más necesitados y con menores oportunidades. Si no se tiene en cuenta este sesgo, los impactos tienden a sobreestimarse ya que los avances entre los que parten de un nivel inferior en la variable a medir, suelen ser más rápidos.

Para poder eliminar el sesgo de selección, la mejor manera de construir un contrafactual "perfecto" es crear un grupo de control cuyas características (observables y no observables), en promedio, sean idénticas al grupo de "tratamiento". Esto se logra mediante la asignación aleatoria entre elegibles. Como he defendido en otra parte (Larrú 2008a), la limitación ética de no beneficiar al grupo de control (supuestamente, pues no hay certeza del beneficio que es precisamente lo que motiva la evaluación) puede salvarse mediante la rotación temporal del tratamiento. Por ejemplo, en el diseño de Banerjee, Cole, Duflo & Linden (2007) en India, las clases que en un año disponían de un refuerzo (T) o de programa informático (*T)* para enseñar matemáticas, no lo tenían al año siguiente, siendo la clase de control quien sí lo disfrutaba, como refleja el Cuadro 3.

Cuadro 3. Ejemplo de diseño de evaluación experimental con asignación aleatorizada.

	Año 1 (2001/02)		Año 2 (2002/03)		Año 3 (2003/04)	
Ciudad: Vadodare	3º	4º	3º	4º	3º	4º
Grupo A	T	C	C	T	C	C
Grupo B	C	T	T	C	C	C
Grupo A1B1	*C*	*C*	*C*	*T*	*C*	*C*
Grupo A2B2	*C*	*C*	*C*	*C*	*C*	*T*
Ciudad: Mumbay						
Grupo C	T	C	C	T	C	C
Grupo D	C	C	T	C	C	C

Nota: Los grupos A,B,C y D reciben como "tratamiento" un alumno de refuerzo para hacer sus deberes; en los del A1B1 y A2B2 el T consistió en aprender matemáticas usando el ordenador.

Esto permite además medir efectos de medio plazo. No obstante, las consideraciones éticas deben tenerse siempre en cuenta, especialmente garantizando el consentimiento informado para participar, tanto como integrante del grupo de tratamiento como de control. Es significativo que en muchos programas de asignación de beneficios, los propios elegibles son conscientes que –bajo la inevitabilidad de la restricción de recursos- la aleatorización (una lotería) es la forma más justa de asignación[14].

Pero el problema de la heterogeneidad aparece cuando se desea conocer no sólo el efecto promedio de una intervención (la diferencia de medias entre el grupo de tratamiento y de control) sino cómo afectó a cada uno de los tratados o a grupos de particular inetrés. Si el hecho de ser tratado está correlacionado con el efecto que pueda tener, la varianza del efecto promedio puede estar sesgada. Deaton (2010) sostiene que el problema de la heterogeneidad no es técnico, sino de teoría. No disponemos de un test econométrico previo que pueda decirnos si la correlación entre "el hecho" de ser tratado y su efecto ex –post es alta.

El problema es análogo a la definición de variables instrumentales (V.I.). Una V.I. se construye para tratar de eliminar el problema de la simultaneidad o causalidad reversible. Por ejemplo, ayudar a una comunidad rural indígena muy pobre tiene correlación con el volumen de la cooperación al desarrollo, ya que ese grupo de población, precisamente por su vulnerabilidad, "atrae" más ayuda que otro. La correlación será directa. Cuanto *más* pobre es la comunidad *más* ayuda habrá que darle. Cuando se mida el efecto ex –post de la ayuda, esperamos que la correlación sea inversa (parámetro negativo): como se ha otorgado *más* ayuda ahora hay *menos* pobreza.

Este problema se magnifica según se amplía el rango de la investigación. Las evidencias empíricas de corte transversal o en datos de panel, no han logrado obtener una conclusión satisfactoria del efecto de la ayuda mundial sobre el crecimiento económico per capita de los países en desarrollo. En los primeros estudios sobre eficacia de la ayuda, no se tomaba en cuenta la simultaneidad ni se usaron V.I. y la

[14] En este sentido, resulta muy iluminador la experiencia de asignación de canalización de agua y saneamiento en Sica-Sica, Bolivia, donde el sorteo se hizo de forma pública entre todos los beneficiarios en la propia plaza del pueblo, precisamente por sugerencia y deseo de transparencia de las autoridades locales más que por los evaluadores. Véase http://vimeo.com/86744573. Hay varios casos más que ilustran el uso de sorteos en experiencias de implementación y evaluación del Banco Mundial disponibles en su revista *World Bank Research Observer*.

mayoría de las evidencias ofrecieron resultados levemente negativos o nulos de la ayuda sobre el crecimiento per capita de los países. Los estudios más recientes que sí consideran la simultaneidad, empleando V.I. geográficas (distancia al ecuador o latitud del país) o políticas (ser ex colonia de un donante) tienden a hallar que el efecto es positivo, aunque no siempre (Rajan & Subramanian 2008; Nowak-Lehman et al. 2012). Las revisiones de la literatura sobre el tema (Larrú 2009; McGillivray et al. 2006; Roodman 2007 a, b y 2008; Tezanos 2010) advierten que la calidad de la V.I. es quizá el aspecto nuclear para poder explicar una diversidad de resultados que últimamente está arrojando más evidencias de efectos positivos (Dovern and Nunnenkamp 2007; Minoui and Reddy 2010; Arndt et al. 2010, 2011; Juselius et al. 2011; Tezanos et al. 2013; Lof et al. 2014) que negativos, aunque el tema sigue abierto incluso tras el uso de meta-análisis (el de Doucouliagos and Paldam [2011] evidencia efecto nulo mientras el de Mekasha and Tarp [2011] positivo)[15].

No conocer la varianza exacta, ni el error estándar del efecto promedio puede ser una limitación importante, sobre todo si es precisamente esa heterogeneidad lo que se quiere saber. Las V.I. no garantizan que la correlación entre la V.I. y los residuos del tratamiento sean ortogonales (su producto escalar sea cero). Pero Banerjee & Duflo (2008) informan sobre posibilidades estadísticas para afrontar el problema como los test de dominancia estocástica (Abadie 2002) o técnicas econométricas sofisticadas (Crump et al. 2008; Imbens & Wooldridge 2009).

Una manera de hacer ver el problema de la heterogeneidad la ofrecen Banerjee & Duflo (2008). Supongamos que los resultados de un tratamiento han sido 2, 3 y -4. En el grupo de control se midieron resultados de 1, 2 y 3. El promedio de T es 0,33 mientras que el de C es 2. ¿Debe recomendarse el tratamiento? El efecto negativo sobre *un* individuo (que tiene un efecto de -4) puede llevar a rechazar el tratamiento, pero eso supone no aprobar un beneficio para *dos* individuos. La simple diferencia de medias entre T y C no es suficiente para llegar a una recomendación operativa. Esta limitación conduce directamente a la necesidad de interpretar los resultados de cada evaluación concreta en el contexto más amplio de una buena teoría causal del cambio.

La importancia de la teoría y la creatividad en los diseños de evaluación.

[15] Es probable que esta incapacidad de consensuar un resultado macro entre los académicos, haya impulsado aún más el desempeño y hasta "moda" de las evaluaciones de impacto que son de naturaleza micro.

Otra de las limitaciones que se atribuyen a las evaluaciones bajo métodos experimentales es que consiguen una buena medida para el impacto, pero que pueden carecer de una teoría consistente para la explicación del *por qué* de esos resultados, un aspecto que precisamente trata de resaltar el diseño evaluativo de la Teoría del Programa (White 2009) o teoría del cambio (Lensink 2014).

La necesidad de una expansión de la teoría en la economía del desarrollo ha sido señalada por varios de los analistas más influyentes: Acemoglu (2010), Basu (2005), Banerjee et al. (2005), Ray (2000 y 2007), entre otros[16].

Si bien pueden existir casos en los que la evaluación se haya centrado en el aspecto de la medición, la afirmación de ausencia de teoría en las evaluaciones de impacto experimentales no puede generalizarse. De hecho, muchas de las evaluaciones más citadas en la literatura sobre el desarrollo se caracterizan por la creatividad y originalidad del diseño evaluativo, precisamente por abordar cuestiones teóricas de alto interés. Por ejemplo, los diseños de Karlan y Zinman (2008, 2009) sobre microcréditos han dado evidencia empírica a cuestiones teóricas concretas como el azar moral y la selección adversa. Precisamente su valor es resaltado por unir a una teoría, una evidencia empírica que la permita seguir evolucionando y concretándose. Karlan y Zinman (2009) revelan que hay azar moral y (menos) selección adversa en la práctica microfinanciera. En concreto entre un 7-16% de la mora se debe a problemas de información asimétrica. En el trabajo de Karlan y Zinman (2008) se plantea la cuestión de si conviene subir el tipo de interés para no crear dependencia en el receptor pobre. Si así fuera, suponemos que el pobre es precio aceptante al tipo de interés al que tome el préstamo y la institución microfinanciera aumenta su beneficio ampliando el alcance del crédito. Los autores muestran que la elasticidad del tomador pobre al precio (interés) del crédito es muy alta. En otras palabras, la curva de demanda de crédito es decreciente y elástica respecto al tipo de interés, pero es el *tamaño del crédito* lo que más explica su madurez, no el tipo de interés. No hay duda de que detrás de su

[16] Precisamente la conferencia anual del Banco Mundial sobre Desarrollo Económico, conocida como *ABCDE Conference,* ha tenido como tema en la edición de 2-3 de Junio de 2014, el papel de la teoría en la economía del desarrollo, con una sesión específica dedicada a la validez de los diseños experimentales. En el discurso de apertura, el economista jefe del Banco Mundial se hizo eco de la importancia de la inferencia (deductiva e inductiva) y el sentido común, así como de la teoría, para complementar las recientes "modas" de los diseños aleatorizados y los trabajos bajo "big data". Ironizó poniendo como ejemplo que el teorema de Pitágoras sea resultado de teoría y que quizá hoy día habría sido objeto de críticas como que no ha sido resultado de un dibujo aleatorio de triángulos, que no haya garantías de su funcionamiento en el cono Sur del planeta o que esté sesgado por los triángulos "mediterráneos" sin ninguna garantía de que el teorema vaya a funcionar en el futuro o en triángulos del grupo de control. Véase Basu (2014).

evaluación hay mucha teoría planteada y el valor de la medición que permitió el RCT ha hecho avanzar el conocimiento tanto empírico como teórico en el campo de las microfinanzas (véase Larrú 2008b para una revisión del tema).

Otro campo de alto interés donde se refleja la dificultad de inferir "leyes universales y necesarias" en economía del desarrollo, son los beneficios que realmente dejan los proyectos de microemprendimiento. El estudio de los incentivos a recibir un microcrédito para abrir o expandir una microempresa está siendo analizado en varias partes del mundo arrojando resultados divergentes. Así, Bandiera et al. (2013) encuentran que un programa en Bangladesh, donde a las mujeres campesinas se las formó y hubo transferencias de activos -además del microcrédito- aumentaron sus ingresos promedio en un 38%. Cull et al. (2008) encuentran una tasa de retorno en beneficios a los micropréstamos en México del 20-30%, pero De Mel et al. (2008) en Sri Lanka en un programa similar, evidencian tan sólo del 5-7%. Una explicación *teórica* de esta diversidad es la de Karlan et al. (2012) y Duflo et al. (2013) que acentúan el pequeño tamaño de los micro-negocios para producir bajo economías de escala o acceder a mejores precios de materias primas por volumen de compras. Así, tras 3-4 años después de recibir el crédito, la elegibilidad era la misma entre los que habían participado y los del grupo de control, aunque el hogar bajo tratamiento obtenía una pequeña cantidad mayor de principal y un mayor plazo de devolución. Sin embargo, no se detectaron cambios estadísticamente significativos ni en el consumo de los hogares, ni en los beneficios de los negocios, ni en la salud, educación o empoderamiento de la mujer. Aunque su evaluación se centró en India, hay evidencia de resultados similares en otros cuatro países (tan diversos como Marruecos, Bosnia, Mongolia y México) lo que cuestiona que la validez externa sea una deficiencia insuperable en el proceso de construcción de conocimiento a partir de evaluaciones aleatorizadas comparables.

La validez externa. El problema de la escalabilidad de resultados.

De alguna manera, la ciencia lo que busca es construir enunciados universales y necesarios, detectar constancias en los resultados y ser capaces de asociarlos a causas comunes. Pero lo que ofrecen las evaluaciones (todas, no sólo las experimentales) son informaciones individuales, notablemente afectadas por los contextos, tanto en sus variables exógenas (clima, geografía, situación sociopolítica) como endógenas a la intervención (la motivación y características de los participantes o el diseño específico

de la intervención, entre otras). Esto hace que la capacidad de las evaluaciones para ofrecer información que se pueda "universalizar" o replicar en otros lugares y circunstancias, pueda ser baja.

Una forma de afrontar esta limitación es la de multiplicar las evaluaciones hasta disponer de tal número de experimentos que podamos crear cierto "paradigma falsable" (en el sentido de inferir inductivamente a partir de las evaluaciones, una teoría que provea de la mejor explicación provisional hasta encontrar otra superior) que sea el que oriente las medidas de política.

Otra opción es considerar que las políticas públicas y programas sociales se implantan en contextos específicos, por lo que el interés "práctico" (político) es precisamente conocer qué funciona en cada contexto y no una "ley universal" que casi nunca se produce en ciencias sociales. Se trataría de crear sistemas de "diagnóstico diferencial particular" (existen enfermos) frente a universales (enfermedades) que quizá sólo se identifican en la física estática, en la ley de la gravedad o en la lógica y matemáticas como únicas ciencias basadas en lenguajes formales construidos precisamente bajo reglas y axiomas[17].

Fue precisamente el caso de la evaluación externa y experimental del programa de transferencias condicionadas en efectivo denominado en 1997 *Progresa* en México y que fue llevada a cabo por IFPRI, lo que se ha utilizado como mejor ejemplo de lo que políticamente la evaluación rigurosa puede incentivar. El diseño inicial del programa fue siendo extendido paulatinamente a más distritos y estados del propio México (pasó a denominarse *Oportunidades* a comienzos de 2002 y *Prospera* desde 2014), desde el ámbito rural al urbano y diversificando las cuantías de las transferencias. En la actualidad, este programa sigue siendo evaluado mientras permanece activo, constituyendo uno de los ejemplos de buenas prácticas evaluativas en ciencia social. Sin necesidad de una garantía de efectos idénticos u homogéneos, numerosos países en América Latina, África o en la ciudad de Nueva York, han ido implantando y ensayando programas de transferencias inspirados en Oportunidades. La adaptación a cada contexto, frente a la copia mimética de lo que funcionó en otro país, es precisamente una clave del éxito de los programas de transferencias. Es más, recientes compilaciones comparativas (Fiszbein & Schady 2007; Cecchini y Madariaga 2011) e informes de

[17] En este sentido, la idea se parece mucho a lo propuesto por Hausmann et al. (2008) con sus "growth diagnosis" en el campo del crecimiento económico.

síntesis (Davies et al. 2012; Baird et al. 2014) más que "cerrar el tema", abren un abanico de enorme riqueza en términos de economía política pues evidencian que pueden funcionar (y no) tanto transferencias condicionadas como no condicionadas, en especie o en efectivo. En el fondo, la riqueza del comportamiento humano es tan grande y las opciones de política tan amplias, que quizá lo que se busque es conocer mejor el efecto "individual" de un programa contextualizado, frente a la construcción de un conocimiento "universal y necesario".

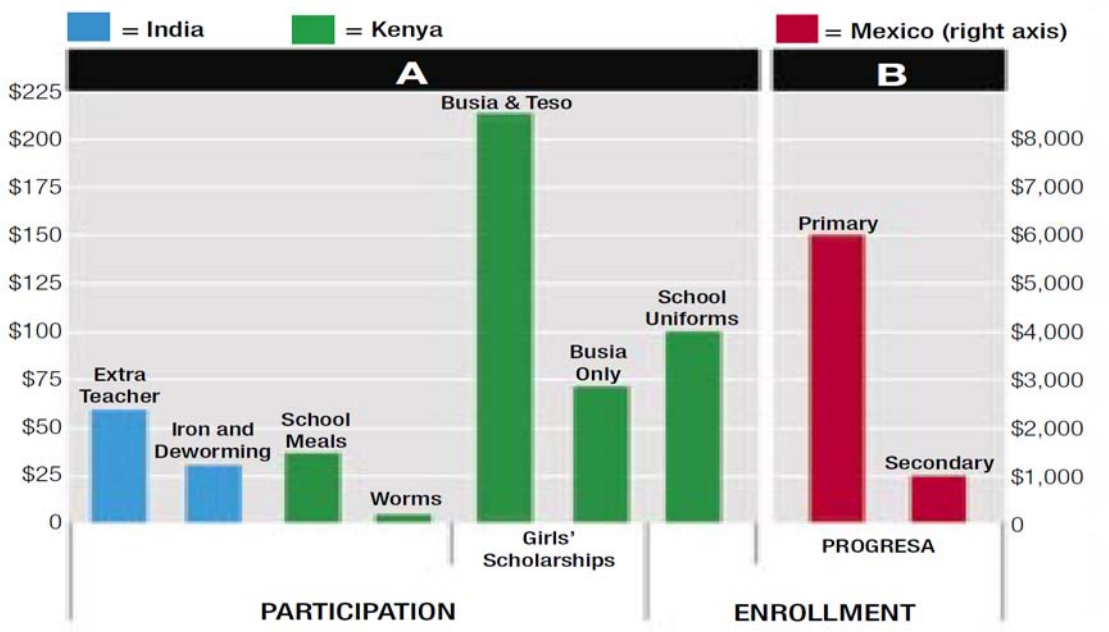
Precisamente otro campo que está siendo expandido y enriquecido a raíz de las evaluaciones de impacto es la "economía del comportamiento" (Bertrand et al. 2004; Ludwig et al. 2011; Mani et al. 2013; Datta & Mullainathan 2014), no sólo en experimentos de laboratorio, sino con ejercicios evaluativos sobre el terreno (Ferraro & Price 2011). La pregunta concreta que abordan estos estudios es si los pobres tienen reacciones, incentivos y conductas propias y diferenciadas de los no pobres. Las evidencias van apuntando que no. Sus preferencias y utilidades, sus resistencias al esfuerzo constante, sus olvidos y tendencias a procrastinar acciones no apetecibles, coinciden con muchas de las nuestras. A todos nos cuesta ponernos a dieta o comer sano, a pesar de apuntarnos al gimnasio y proponernos cada año cuidar mejor nuestro cuerpo. Lo que pasa es que tendemos a juzgar como "no racional" que una familia pobre tenga televisión o el cabeza de familia beba demasiado para evadirse cuando sus –para nosotros- demasiados hijos presentan desnutrición y no acuden con regularidad a la escuela, que por otra parte puede que esté muy alejada y el profesor ni acuda o simplemente esté borracho y por eso haya padres que prefieran pagar una escuela privada aun siendo muy pobres (véase el magnífico desarrollo de este punto en Tooley 2009).

En el fondo, el cuerpo de literatura sobre *behavioral economics* trata de avanzar en la compleja cuestión –hasta ahora sin resolver- de si existen o no en la práctica las "trampas del desarrollo" o círculos viciosos de la pobreza (Bowles et al. 2006; Kraay & Raddatz 2007). Esta cuestión es la que está en el fondo de la influyente obra de Banerjee y Duflo (2011) y sobre la que una reciente revisión de la literatura y ejercicio métrico (McKay & Perge 2011; Kraay & McKenzie 2014) ha llegado a la conclusión (parcial) de que, o no existen o, si se dan, lo hacen en una escala y entorno tan pequeño que no puede considerarse una barrera generalizable para los países que aun arrastran enormes bolsas de pobreza. Precisamente este paso de considerar las problemáticas de

la pobreza y el subdesarrollo desde una óptica macroeconómica y algo "abstracta", a un enfoque microeconómico y más concreto (como mucho sectorial), es otra de las características propias de lo que se ha denominado "la vuelta de la economía del desarrollo a ser influyente" (Leonhardt 2008).

En resumen, muchas de las críticas que se hacen a los RCT[18] se matizan mucho si se tiene en cuenta dos aspectos esenciales en la evaluación: primero *cuál es la pregunta exacta a la que se desea responder* y segundo, *cuál es el diseño mejor para poder construir una respuesta satisfactoria*. El diseño experimental no es el único. No es perfecto. No es la única manera de hacer ciencia rigurosa. Pero debe reconocerse que en muchas ocasiones es el más potente y capaz de responder a cuestiones de alto calado científico y político. Es cierto que se han puesto de moda y puede caerse en la tentación de menospreciar a todos los demás, señalándolos como "no científicos" o "evidencia blanda". Eso tampoco debería ocurrir, como no fue deseable la situación antes del movimiento de "rescate" de los RCT, de considerar que la evaluación de impacto era imposible en ciencias sociales. Emplear los RCT en preguntas muy concretas y ámbitos paralelos está permitiendo, quizá por primera vez, hacer estudios de eficiencia rigurosos ya que el método evaluativo ha sido el mismo (por ejemplo en los casos de eficiencia para afrontar el absentismo escolar, Duflo 2008). No puede negarse que hasta ahora no había sido posible conocer que, el coste por año adicional de educación primaria, es mucho menor con programas de desparasitación que con becas o transferencias condicionadas. Al medir el impacto, se ha podido comparar con rigor y tomar decisiones políticas eficientes para una misma cuestión aparentemente sencilla: ¿cómo hacer que los niños y niñas acudan más a clase? Desparasitando cuesta 3,5 dólares anuales y dar becas a niñas en Kenia costó más de 200 dólares por año escolar adicional. Las transferencias condicionadas en México, para un año adicional en secundaria, se elevan a 1.000 dólares y 6.000 dólares en primaria, dada la alta propensión a acudir ya a la escuela en esa etapa escolar. Las diferencias presupuestarias hablan por sí mismas como muestra el Cuadro 4.

[18] Harrison (2014) destaca tres: limitarse a analizar observables; limitarse a medir el efecto promedio; limitarse al análisis de equilibrio parcial. Propone profundizar en el análisis coste-eficiente de bienestar profundizando en "estándares", rigor y pertinencia, e importancia del poder y la política.



Cuadro 4. Coste por año adicional de escolarización.

Procesos comparativos similares se están pudiendo hacer gracias a evaluaciones de impacto entre programas de lucha contra el SIDA (Kremer 2008) o la elasticidad precio de servicios básicos de salud como el acceso a mosquiteras (Cohen & Dupas 2007; Tarozzi et al. 2014). Probablemente la recopilación de Dhaliwall et al. (2011) sea una de las pocas que hayan abordado de una manera rigurosa el análisis de coste-eficiencia en intervenciones de educación (en concreto cómo lograr un año adicional de escolarización infantil) sobre la base empírica que han aportado las evaluaciones de impacto. Quizá pueda sorprender –por su relativa simpleza- que la intervención más coste-efectiva, resultó ser la explicación a los padres del valor de mayor escolarización de sus hijos en Madagascar (20,7 años por cada 100 dólares invertidos).

En suma, las evaluaciones incentivan a acotar la dimensión de las preguntas y pueden ofrecer *causalidades* más creíbles, pero encontramos un *trade-off* entre causalidad (micro) y agregación, validez externa o generalización. Cuanto más acotamos la hipótesis y la muestra, más manejable, creíble y útil es la relación causal que se pretende afirmar, pero menor es la confianza de que la extrapolación de otra intervención de cooperación al desarrollo, produzca el mismo efecto en otro receptor. Por poner un ejemplo del campo de la medicina: no podemos afirmar que un conjunto acotado de síntomas, producen siempre la misma enfermedad, que tratada de la misma

forma sobre todos los pacientes, producen siempre su curación. Las ciencias sociales no enuncian nunca ese tipo de leyes.

Por último, el método experimental está siendo un método que logra aunar los conocimientos más especializados obtenidos por los investigadores con los intereses concretos de tomadores de decisión y con ONGs sobre el terreno que les ayudan. Por ejemplo, en el caso del estudio seminal de la desparasitación de Miguel y Kremer (2004), la pregunta no fue cómo eliminar la anemia de los niños de Kenia. Partió de una constatación económica importante. Dadas las externalidades que suponen un tratamiento desparasitador, es mejor desempeñar una campaña masiva de administración de medicamentos. Para conocer su efecto, se empezó con una evaluación aleatorizada sobre 75 colegios. Cuando se supo científicamente el impacto gracias al RCT, el programa pudo elevarse de escala y se ha podido comparar su coste-efectividad respecto a otros programas de lucha contra el absentismo escolar (mejoró un 25% y sólo cuesta 3,5 dólares por año adicional de escolaridad infantil) mucho más caros como dar comidas gratuitas (36 dólares) o regalar los uniformes (99 dólares). Además ha permitido comprender que la elasticidad renta-precio entre los participantes del programa es muy alta (rígida). En cuanto dejó de subvencionarse y se cobró el coste de recuperación del medicamento (30 céntimos), la participación descendió en un 80%. Estamos teniendo evidencias semejantes de programas sociales, educativos y sanitarios, muy sensibles al precio (Holla y Kremer 2008). Este conocimiento está siendo posible gracias a evaluaciones bajo diseños experimentales que permiten su comparación. Una consecuencia directa de este impacto está siendo el cuestionamiento de la sostenibilidad de estos programas si se dejan de subvencionar públicamente (Kremer y Miguel 2007). Sin las evaluaciones experimentales, probablemente seguiríamos creyendo que el criterio de sostenibilidad se logra cuando hacemos al usuario capaz de hacerse cargo de su coste de forma privada. Dado el carácter de bien público que tiene la desparasitación (por sus externalidades positivas) esto no tiene ya por qué ser así y el consejo que puede recibir un ministro de sanidad de un país en vías de desarrollo como Kenia, es que debe contar con un "coste fijo" para subvencionar el tratamiento de parásitos o la entrega de mosquiteras (la elasticidad-precio a la compra y uso de este método contra la malaria se ha revelado también muy alto como muestran Cohen y Dupas 2007).

4. Conclusiones.

¿Quién demanda conocimiento evaluativo hoy en día? ¿Sirven realmente para algo las evaluaciones de impacto? ¿Influyen en la toma de decisiones o son sólo una "moda" pasajera exclusiva del ámbito académico?

Puede llamar la atención que análisis de demanda de conocimiento dentro del Banco Mundial como los de Ravallion (2011) o Doemeland y Trevino (2014) evidencien que, en concreto, el 87% de los informes-país de dicha institución no se citen nunca, que el 31% nunca se descarguen de Internet o que sólo el 13% de los *Policy Reports* superen las 250 descargas. Son los informes más caros, complejos, muti-sectoriales, de diagnóstico sobre un problema importante ("*core diagnosis*") y que atañen a los países de renta media muy poblados los que, con mayor probabilidad, son descargados. Pero las descargas aumentan precisamente cuando se han realizado esfuerzos por una difusión interna amplia, debatida y coordinada por el departamento de investigación.

En efecto, como puso de manifiesto Feinstein (2002), el uso no queda garantizado porque un informe esté disponible en internet, sino por un manejo adecuado de la credibilidad, oportunidad y diseminación de dicho informe. La diseminación implica buenos soportes y medios de comunicación, pero sin el rigor y la oportunidad política de la cuestión evaluada, su influencia queda muy mermada como han puesto de relieve estudios sobre esta cuestión (Court et al. 2004; ODI 2004; World Bank 2004; Ramalingam 2011).

Como hemos expuesto en este trabajo, la pregunta de si la cooperación internacional al desarrollo reduce la pobreza o no de forma empírica puede resultar demasiado ambiciosa para obtener una respuesta concluyente. Preguntas del alcance como si es racional pensar que dar ayuda puede causar menor pobreza en sus destinatarios y diseñar la política económica normativa óptima, que maximice la productividad de la ayuda mundial en términos de reducción de la pobreza global, parecen excesivas para el ámbito evaluativo aunque sean plenamente científicas. Y esta cuestión no la puede responder, como hemos visto, un cálculo ni un modelo, pues los datos "observables" de los que disponemos cuando nos movemos en los niveles agregados de la ayuda y la pobreza, contienen tantos supuestos y márgenes de error que no ofrecen respuestas fiables (Larrú 2009).

En definitiva, partir de premisas con tantos supuestos, dificultad práctica y márgenes de error en su medición como cuánta pobreza hay y cuánta ayuda llega realmente de forma eficaz a los pobres, nos debería aconsejar no seguir intentando llegar a conclusiones válidas, ni por orden empírico, ni por orden de razón "universal y necesaria", más bien voluntarista y probablemente con intereses creados a priori por sus defensores. Pueden ponerse ejemplos similares de preguntas pseudo-evaluativas excesivamente ambiciosas: evaluar si la Cooperación Española contribuye al logro o no de los Objetivos del Milenio; evaluar el "impacto" del Plan Director o de un Plan General de Cooperación de una Comunidad Autónoma o municipio, son preguntas desenfocadas o que deben acotarse tras un proceso de negociación de términos de referencia posibles, útiles y que respondan –en principio- a una sola pregunta central: una hipótesis nuclear sobre la que el promotor de la evaluación tenga márgenes de maniobra y acción sobre ella. Lograr acotar la colección de preguntas evaluativas y concretar con precisión el objetivo de cada evaluación es una tarea que bien puede incluirse dentro de las habilidades y misiones que debería tener un evaluador dado su amplio conocimiento de metodologías.

Los impactos a través de diseños experimentales están aportando novedad, creatividad y retos a las teorías microeconómicas que, quizá hasta ahora se han construido sobre bases antropológicas muy limitadas como la neoclásica. La pobreza hace que las personas se comporten de forma "irracional", si la racionalidad es la neoclásica que sólo se orienta por la maximización de un beneficio utilitario tipo benthamita. Banerjee y Duflo (2011) recogen numerosos ejemplos y testimonios que retan el saber "convencional" sin que aun podamos tener seguridad de si existen o no mecanismos circulares o trampas de pobreza.

En suma, si la comunidad académica y evaluativa es capaz de realizar preguntas ajustadas que puedan responderse de forma bastante exacta bajo métodos experimentales, los decisores políticos podrán recibir respuestas "científicas" (evidencias de alto rigor) ante programas que pueden lograr beneficios bajo hipótesis y diseños alternativos. Es deseable que esta opción metodológica sea conocida y apoyada por las Administraciones Públicas (central y descentralizada) que financian las intervenciones en cooperación internacional para el desarrollo.

5. Referencias Bibliográficas

Abadie, A. (2002). Bootstrap Test for Distributional Treatment Effects in Instrumental Variables Models. *Journal of the American Statistical Association 97 (457)*, 284-292.

Acemoglu, D. (2010). Theory, General Equilibrium and Political Economy in Development Economics. *Journal of Economic Perspectives 24(2)*, 17-32.

American Evaluation Association (2003). Response To US Department of Education Notice of Proposed Priority. Federal Register RIN 1890-ZA00, November 4, 2003 'Scientifically Based Evaluation Methods'.

Arndt, Ch. Jones S., & Tarp, F. (2010). Aid, Growth, and Development. Have We Come Full Circle?, *Journal of Globalization and Development 1(2)*, 1-26.

Arndt, Ch. Jones S., & Tarp, F (2011). Aid Effectiveness. Opening the Black Box. UNU-WIDER Working Paper 44.

Baird, S. Ferreira, F.H.G. Özler, B. & Woolcock, M. (2014). Conditional, Unconditional and Everything in Between: A Systematic Review of the Effects of Cash Transfer Programmes on Schooling Outcomes. *Journal of Development Effectiveness 6(1)*, 1-43.

Bandiera, O. Burgess, R. Das, N. Gulesci, S. Rasul, I. & Sulaiman, M. (2013). Can Basic Entrepreurship Transform the Economic Lives of the Poor? IZA Working Paper 7386.

Banerjee, A.V. (2006). The Best Argument for the Experimental Approach Is that It Spurs Innovation, *Boston Review 31 (4)*.

Banerjee, A.V. et al. (2005). New Directions in Development Economics: Theory or Empirics. BREAD Working Paper 106, A Symposium in *Economic and Political Weekly*.

Banerjee, A.V. Cole, S. Duflo, E. & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics 122 (3),* 1235-1264.

Banerjee, A.V. & Duflo, E. (2011). *Poor Economics. A Radical Rethinking of the Way to Fight Global Poverty*. New York. Public Affairs.

Basu, K. (2005). The New Empirical Development Economics: Remarks on its Philosophical Foundations. *Economic and Political Weekly, October 1, 40(40),* 4336-39.

Basu, K. (2005). The New Empirical Development Economics: Remarks on its Philosophical Foundations. in Kanbur, R. (Ed.) New Directions in Development Economics: Theory or Empirics? A symposium in *Economic and Political Weekly*, typescript, August.

Basu, K. (2014). Development Economics and Method: A Quarter Century of ABCDE. blog post on *Let`s Talk Development 06/04/2014.*

Bertrand, M. Mullainathan, S. & Shafir, E. (2004). A Behavioral Economics View of Poverty. *American Economic Review, 94 (2),* 419-423

Blattman, C. (2008). Impact Evaluation 2.0. Presentation to the DFID, London.

Bowles, S. Durlauf, S. & Hoff, K. (eds.) (2006). *Poverty Traps*. New York. Russell Sage Foundation.

Camfield, L. & Duvendack, M. (2014). Impact Evaluation – Are We'Off the Gold Standard? *European Journal of Development Research 26(1),* 1-11.

Cecchini, S. & Madariaga, A. (2011). *Programas de transferencias condicionadas. Balance de la experiencia reciente en América Latina y el Caribe.* Santiago de Chile. CEPAL y ASDI.

Center for Global Development (2006). When Will We Ever Learn? Improving Lives through Impact Evaluation. Washington, DC. Report of the Evaluation Gap Working Group.

Cohen, J. & Dupas, P. (2007). Free Distribution or Cost-Sharing ? Evidence from a Randomized Malaria Prevention Experiment. Brookings Global Economy and Development Working Paper 11.

Court, J., Hovland, I. & Young, J. (2004). Bridging Research and Policy in International Development: Evidence and the Change Process. London. ITDG.

Crump, R. Hotz, J. Imbens, G. & Mitnik, O. (2008). Nonparametric Tests for Treatment Effect Heterogeneity. *Review of Economics and Statistics 90(3)*, 389-405.

Cull, R. McKenzie, D. & Woodruff, Ch. (2008). Experimental Evidence on Returns to Capital and Access to Finance in Mexico. *World Bank Economic Review 22(3)*, 457-482.

Datta, S. & Mullainathan, S. (2014). Behavioral Design: A New Approach to Development Policy. *Review of Income and Wealth 60, (1)*, 7-35.

Davis, B. Gaarder, M. Handa, S. & Yablonski, J. (2012). Evaluating the impact of cash transfer programmes in Sub-Saharan Africa : An Introduction to the Special Issue. *Journal of Development Effectiveness 4(1)*, 1-8.

De Mel, S. McKenzie, D. & Woodruff, CH. (2008). Returns to Capital in Microenterprises: Evidence from a Field Experiment. *Quarterly Journal of Economics 123(4)*, 1329-1372.

Deaton, A. (2006). Evidence-based aid must not become the latest in a long string of development fads. *Boston Review 31* (4).

Deaton, A. (2009). Instruments of Development Randomization in the Tropics and the Search fro the Elusive Keys to Economic development. NBER WP 14690.

Deaton, A. (2010). Instruments, Randomization and Learning about Development. *Journal of Economic Literature 48(2)*, 424-455.

Dhaliwal, I; Duflo, E. Glennerster, R. & Tulloch, C. (2011). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General

Framework with Applications for Education. J-PAL Working Paper.

Doucouliagos, H. & Paldam, M. (2006). The Aid Effectiveness Literature. The Sad Result of 40 Years of Research. University of Aarhus, Department of Economics Working Paper 18/4-2006.

Doucouliagos, H. & Paldam, M. (2011). The Ineffectiveness of Development Aid on Growth: An Update Covering Four Years of Research. *European Journal of Political Economy 27(2),* 399-404.

Dovern, J. & Nunnenkamp, P. (2007). Aid and Growth Accelerations: An Alternative Approach to Assessing the Effectiveness of Aid. *Kyklos* 60(3), 359-383.

Duflo, E. (2004). Scaling up and Evaluation. in Bourguignon, F. & Pleskovic, B. (Eds.) *Accelerating Development*, Washington. World Bank & Oxford Univ. Press. pp. 342-367.

Duflo, E. (2005). Field Experiments in Development Economics, paper prepared for the World Congress of the Econometric Society, December. in Blundell, R. Newey, W. & Persson, T. (Eds.) *Advances in Economic Theory and Econometrics*, Cambridge Univ. Press. vol 2(42).

Duflo, E. (2008). La evaluación de las intervenciones educativas: evidencia a partir de experimentos aleatorizados. en Montalvo, J.G. (ed.) *El análisis experimental de la ayuda al desarrollo. La evaluación de lo que funciona y lo que no funciona*". Madrid. Fundación BBVA. Cap 3, 75-102.

Duflo, E. & Kremer, M. (2005). Use of Randomization in the Evaluation of Development Effectiveness. in Pitman, G. Feinstein, O. & Ingram, G. (eds) *Evaluating Development Effectiveness.* World Bank Series on Evaluation and Development. Vol.7. Transaction Publishers. New Brunswick.

Duflo, E. Banerjee, A. Glennerster, R. & Kinnan, C. (2013). The miracle of microfinance? Evidence from a randomized evaluation. NBER Working Paper 18950.

European Evaluation Society (2007). EES Statement: The Importance of a Methodology Diverse Approach to Impact Evaluation – Specifally with Respect to Development Aid and Development Interventions.

Feinstein, O. (2002) Use of Evaluations and the Evaluation of their Use. *Evaluation 8(4)*, 433-439.

Feinstein, O. (2012). Informe de Evaluación del Programa del Fondo Español para la Evaluación de Impacto (SIEF). Informe Final. 2 de junio.

Ferraro, P. & Price, M. (2011). Using Non-Pecuniary Strategies to Influence Behavior: Evidence from a Large Scale Field Experiment. NBER Working Paper 17189.

Fiszbein, A. & Schady, N. (eds.) (2009). *Conditional Cash Transfers. Reducing Present and Future Poverty*. Washington. The World Bank.

Foss Hansen, H. & Rieper, O. (2009). The Evidence Movement. The Development and Consequences of Methodologies in Review Practices. *Evaluation 15(2)*, 141-163.

Forss, K. & Bandstein, S. (2008). "Evidence-based Evaluation of Development Cooperation: Possible? Feasible? Desirable?", NONIE Working Paper 8.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*, 424-438.

Guijt, I. & Roche, Ch. (2014). Does Impact Evaluation in Development Matter? Well, It Depends What It's For! *European Journal of Development Research 26(1)*, 46-54.

Karlan, D. & Zinman, J. (2008). Credit Elasticities in Less-Developed Economies : Implications for Microfinance. *American Economic Review 98 (3)*, 1040-68.

Karlan, D. & Zinman, J. (2009). Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment. *Econometrica 77 (6)*, 1993-2008.

Kraay, A. & McKenzie, D. (2014). Do Poverty Traps Exist? Assessing the Evidence. *Journal of Economic Perspectives 28(3),* 127-148

Harrison, G.W. (2014). Impact Evaluation and Welfare Evaluation. *European Journal of Development Research 26(1),* 39-45.

Hausmann, R. RODRIK, D. & VELASCO, A. (2008). Growth Diagnostics. in Stiglitz, J. & Serra, N. (Eds.) *The Washington Consensus Reconsidered: Towards a New Global Governance.* New York. Oxford University Press.

Heckman, J. (2008). Econometric Causality. NBER Working Paper 13934.

Holla, A. & Kremer, M. (2008) Pricing and Access: Lessons from Randomized Evaluations in Education and Health. paper for Brookings Global Economy and Development Conference: "What Works in Development? Thinking Big and Thinking Small", 29-30 may.

ILO. (2002). Extending Social Protection in Health trough Community Based Health Organizations: Evidence and Challenges. Geneva. Discussion Paper. Universitas Programme.

Imbens, G. & Wooldridge, J. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, 47(1),* 5-86.

J-PAL. (2005). Fighting Poverty: What Works? J-PAL Policy Brief 1.

Juselius, K. Framroze, N. & Tarp, F. (2011). The Long-Run Impact of Foreign Aid in 36 African Countries. UNU-WIDER Working Paper 51.

Karlan, D. Knight, R. & Udry, Ch. (2012). Hoping to Win, Expected to Lose: Theory and Lessons on Micro Enterprise Development. NBER Working Paper 18325.

Kraay, A. & Raddatz, C. (2007). Poverty Traps, Aid, and Growth. *Journal of Development Economics 82(2),* 315-347.

Kremer, M. & Miguel, E. (2007). The Illusion of Sustainability. *Quarterly Journal of Economics 122(3),* 1007-1065.

Kremer, M. (2008). Cambios en los hábitos y los sistemas sanitarios: algunas evidencias a partir de evaluaciones aleatorizadas, en Montalvo, J.G. (Ed.) *El análisis experimental de la ayuda al desarrollo. La evaluación de lo que funciona y lo que no funciona*". Madrid. Fundación BBVA. Cap 2, 49-74.

Larrú, J.M. (2008a). La Evaluación de Impacto: ¿Qué es, Cómo se mide y qué está aportando en la Cooperación al Desarrollo. en Larrú, J.M. (coord.) *Evaluación en la Cooperación para el Desarrollo.* Madrid. *Colección Escuela Diplomática 12*, 109-133.

Larrú, J.M. (2008b). Las evaluaciones de impacto con asignación aleatoria y los microcréditos. *Revista de Economía Mundial 19*, 34-62.

Larrú, J.M. (2009). *La ayuda al desarrollo, ¿reduce la pobreza? Eficacia y evaluación de la cooperación para el desarrollo.* Madrid. Biblioteca Nueva.

Larrú, J.M. (2010). Algunas cuestiones conceptuales y metodológicas en torno a la evaluación de impacto. *Revista E-valuación 11*, 20-31.

Larrú, J.M. (2011). Evaluaciones en la cooperación para el desarrollo: promesas y amenazas. Nombres Propios. Fundación CEALCI.

Larrú, J.M. (2012). Las brechas de la evaluación en la cooperación española al desarrollo. *Revista Española del Tercer Sector 22*, 93-118.

Larrú, J.M. & Méndez, M. (2012). La utilidad de las evaluaciones en las ONGD españolas: un estudio basado en la convocatoria de convenios AECID 2006 y 2007. *Revista de Fomento Social 67(267)*, 449-485.

Lensink, R. (2014). What Can We Learn from Impact Evaluation? *European Journal of Development Research 26(1)*, 12-17.

Leonhardt, D. (2008). Making Economics Relevant Again, New York Times, Feb, 20.

Levine, R. (2004). *Millions Saved: Proven Successes in Global Health.* Washington. What Works Working Group. Center for Global Development.

Lloyd, T. McGillivray, M. Morrissey, O. & Osei, R. (2000). Does Aid Create Trade? An Investigation for European Donors and African Recipients. *European Journal of Development Research 12 (1),* 107-123.

Lof, M. Mekasha, T.J. & Tarp, F. (2014). Aid and Income: Another Time-series Perspective. *World Development* (in press), http://dx.doi.org/10.1016/j.worlddev.2013.12.015

Ludwig, J. Kling, J. & Mullainathan, S. (2011). Mechanism Experiments and Policy Evaluations. NBER Working Paper 17062.

MAEC. (2012). *La Evaluación en la Cooperación Española. Informe Anual 2010*. Madrid. MAEC-SECIPI Secretaría General de Cooperación Internacional para el Desarrollo- División de Evaluación y Gestión del Conocimiento.

MAEC. (2013a). *Política de Evaluación de la Cooperación Española*. Madrid. MAEC-SECIPI Secretaría General de Cooperación Internacional para el Desarrollo- División de Evaluación y Gestión del Conocimiento.

MAEC. (2013b). *Plan Bienal de Evaluaciones 2013-2014*. Madrid. MAEC-SECIPI Secretaría General de Cooperación Internacional para el Desarrollo- División de Evaluación y Gestión del Conocimiento.

Mani, A. Mullainathan, S. Shafir, E. & Zhao, J. (2013). Poverty Impedes Cognitive Function. *Science 34 (6149),* 976-980.

McGillivray, M. Feeny, S. Hermes, N. & Lensink, R. (2006). Controversies Over the Impact of Development Aid: It Works; It Doesn´t; It Can, But That Depends. *Journal of International Development 18 (7),* 1031-1050.

McKay, A. & Perge, E. (2011). How Strong Is the Evidence for the Existence of Poverty Traps? A Multi Country Assessment, University of Sussex Economics Department Working Paper Series 25.

Mekasha, T. J. & Tarp, F. (2011). Aid and Growth. What Meta-Analysis Reveals. UNU-WIDER Working Paper 22.

Miguel, E. & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, LXXII* 159-217.

Minoiu, C. & Reddy, S. (2010). Development Aid and Economic Growth: A Positive Long-Run Relation. *The Quarterly Review of Economics and Finance 50 (1),* 27-39.

Nowak-Lehmann, D. F. Dreher, A. Herzer, D. Klasen, S. & Martínez-Zarzoso, I. (2012). Does Foreign Aid Really Raise Per-capita Income? A Time Series Perspective. *Canadian Journal of Economics 45 (1),* 288-313.

ODI. (2004). Bridging Research and Policy in International Development An analytical and practical Framework. Overseas Development Institute Briefing Paper.

Picciotto, R. (2014). Is Impact Evaluation Evaluation? *European Journal of Development Research 26(1),* 31-38.

Rajan, R. & Subramanian, A. (2008). Aid and Growth: What Does the Cross-Country Evidence Really Show? *Review of Economics and Statistics 90 (4),* 643-665.

Ramalingam, B. (2011). Learning How to Learn: Eight Lessons for Impact Evaluations that Make a Difference. ODI Background Note. April.

Rao, V. (2008). El valor de la evaluación interdisciplinar: el análisis de programas de desarrollo basados en la comunidad. en Montalvo, J.G. (Ed.) *El análisis experimental de la ayuda al desarrollo. La evaluación de lo que funciona y lo que no funciona*". Madrid. Fundación BBVA. Cap 5, 129-145.

Ravallion, M. (2011). Knowledgeable Bankers? The Demand for Research in World Bank Operations. World Bank Policy Research Working Paper 5892.

Ray, D. (2000). What's New in Development Economics? *The American Economist 44,* 3-16.

Ray, D. (2007). Development Economics. Prepared for the New Palgrave Dictionary of Economics, edited by Lawrence Blume and Steven Durlauf.

Rodrik, D. (2008). The New Development Economics: We Shall Experiment, But How Shall We Learn? Paper for Brookings Global Economy and Development Conference: "What Works in Development? Thinking Big and Thinking Small", 29-30 may.

Roodman, D. (2007a). The Anarchy of Numbers: Aid, Development, and Cross-country Empirics. World Bank Economic Review 21 (2), 255-277.

Roodman, D. (2007b) Macro Aid Effectiveness Research: A Guide for the Perplexed. Center for Global Development Working Paper 134.

Roodman, D. (2008). Through the Looking Glass, and What OLS Found There: On Growth, Foreign Aid, and reverse Causality. Center for Global Development Working Paper 137.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research, *Journal of MultiDisciplinary Evaluation 5 (9),* 11-24.

Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L. & Yoong, J. (2014). Micro-loans, Insecticide-Treated Bednets and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India. *American Economic Review 104(7)*, 1909-1941.

Tezanos, S. (2010). Ayuda y Crecimiento: una relación en disputa. *Revista de Economía Mundial 26,* 237-259.

Tezanos, S., Guijarro, M. & Quiñones, A. (2013). Inequality, Aid and Growth: Macroeconomic Impact of Aid Grants and Loans in Latin America and the Caribbean. *Journal of Applied Economics XVI (1),* 157-182.

Tooley, J. (2009). *The Beautiful Tree. A Personal Journey Into How the World's Poorest People Are Educating Themselves*. Washington. Cato Institute.

Udry, Ch. (2011). Esther Duflo: 2010 John Bates Clark Medalist. *Journal of Economic Perspectives 25(3),* 197-216.

Victoria, C.G. (1995). A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992-1993, Evaluation & Research Working Paper Series Nº 3. UNICEF. New York.

White, H. (2009). Theory-Based Impact Evaluation: Principles and Practice. *Journal of Development Effectiveness 1(3)*, 271-284.

White, H. (2014). Current Challenges in Impact Evaluation. *European Journal of Development Research 26(1)*, 18-30.

World Bank. (2004). Influential Evaluations: Evaluations that Improved Performance and Impacts of Development Programs. Washington. The World Bank.