Impact Assessment and Evaluation: What it is it, how can it be measured and what it is adding to the development of international co-operation.

Larru, Jose Maria

December 2007

LA EVALUACIÓN DE IMPACTO: QUÉ ES, CÓMO SE MIDE Y QUÉ ESTÁ APORTANDO EN LA COOPERACIÓN PARA EL DESARROLLO.

José María Larrú*

Artículo incluido en **LARRÚ, J.M. (coord.) (2007)** *Evaluación en la Cooperación para el Desarrollo.* Colección Escuela Diplomática Nº12. Madrid. pp.109-133.

Resumen:

La evaluación de impacto de intervenciones de cooperación para el desarrollo permite generar un conocimiento preciso sobre lo que funciona y lo que no en la ayuda. El impacto resuelve la inobservancia del contrafactual y el problema de la atribución. La mejor manera de hacerlo es mediante un diseño experimental o evaluación *randomizada.* En el trabajo se expone cómo se realiza dicho diseño, cuándo conviene hacerlo y cuándo no. Se discuten sus ventajas y limitaciones frente a diseños alternativos como los no experimentales y los participativos. Se ilustra su capacidad de generar información rigurosa y cuantitativa con ejemplos del sector . Por último, se exponen recomendaciones para la política española de desarrollo.

Palabras clave: atribución, ayuda al desarrollo, contrafactual, evaluación, impacto.

* Doctor en economía y profesor del Departamento de Economía General de la Universidad CEU-San Pablo de Madrid. Investigador del Centro de Estudios de Cooperación para el Desarrollo (CECOD). Agradezco los útiles comentarios a versiones preliminares del trabajo a Miguel Almunia-Candela; Osvaldo Feinstein; Dan Levy y los participantes en la IX Reunión de Economía Mundial en Madrid, siendo de aplicación las habituales atribuciones de los errores exclusivamente al autor.

1. INTRODUCCIÓN

Supongamos que un responsable político o técnico recién llegado a un nuevo puesto directivo en Cooperación para el Desarrollo formulara la siguiente pregunta a los diferentes agentes que practican la ayuda desde hace tiempo: "¿podría decirme exactamente qué funciona y qué no en la ayuda para reducir la pobreza?" Es muy probable que las respuestas se orientaran a relatar las muchas actividades que gobiernos, organismos internacionales y ONGD llevan años realizando en el campo de la seguridad alimentaria, la educación básica, el acceso a servicios básicos como el agua y el saneamiento, la importancia de factores transversales como el enfoque de género y el medio ambiente y un largo etc. Pero no se estaría respondiendo a la pregunta. Siguiendo el diálogo imaginado, el político podría replicar: "no le pregunto qué sabe hacer usted o a qué se ha dedicado durante años su organización, sino que quiero certezas de a qué debo destinar el escaso dinero de cooperación para el desarrollo, estando seguro de que tiene un efecto probado en la reducción de la pobreza. Quiero saber lo que funciona, lo que no y cuánto cuesta". De alguna manera, este político imaginado no hace sino preguntar al mundo de la cooperación por los impactos.

El impacto es uno de los cinco criterios "clásicos" en las evaluaciones de las intervenciones de cooperación para el desarrollo, sean proyectos, programas, apoyos presupuestarios o políticas. El Comité de Ayuda al Desarrollo (CAD) de la OCDE que agrupa a los principales donantes de ayuda incluye en sus "Principios para una ayuda eficaz (CAD 1991 y 1998) el impacto junto a la pertinencia, eficacia, eficiencia y sostenibilidad de las intervenciones. Asimismo, la Metodología de Evaluación de la Cooperación Española (MAE 1998 y 2001) recoge estos mismos cinco criterios evaluativos y los Informes de Evaluación hechos públicos desde entonces contienen un diagnóstico o juicio sobre el impacto. Pero, ¿qué es exactamente el impacto? ¿Qué sistema de información es necesario establecer para poder evaluarlo? ¿Cómo puede medirse? Este artículo pretende arrojar luz sobre estas tres cuestiones.

Muchos de los contenidos que en los informes de evaluación aparecen como juicio del impacto, en realidad son sólo impresiones subjetivas o –en el mejor de los casos- reconocimiento de que no hay información suficiente para poder realmente pronunciarse sobre el impacto de una intervención. Los principales problemas para juzgar el impacto son la atribución y la agregación[1]. La *atribución* consiste en asignar (atribuir) de forma rigurosa los

[1] Véase la Metodología de Evaluación de la Cooperación Española II (MAE 2001:48) y Roche (2004) pp.105-114 para la atribución y pp.114-123 para la agregación, corroboración y triangulación de los resultados y la

cambios detectados (observables y no observables) a la intervención que se juzga / evalúa y no a otras intervenciones que pueden estar siendo ejecutadas sobre la misma población o en la misma zona. El problema de la *agregación* consiste en resolver cómo poder sumar efectos que se estiman a diferentes niveles y con frecuencia utilizando diferentes unidades de medida. Por ejemplo, un proyecto de capacitación productiva puede mejorar los ingresos de una familia (medidos en unidades monetarias locales); esos mayores ingresos suponen una mejora en la composición de la dieta (medida en calorías o proteínas consumidas) que se traduce en menor morbilidad (medida en número de días de enfermedad), mayor peso de la población infantil (en kilogramos) y menor absentismo escolar (medido en número de días). Cuando todos los efectos se dirigen en una misma dirección (aumento de bienestar) la agregación puede ser menos problemática que cuando se identifican efectos contrarios (que no contradictorios)[2]. En determinados contextos, ambos pueden solucionarse de manera óptima mediante diseños experimentales que, a través del establecimiento de grupos de tratamiento y control configurados de forma aleatoria, permiten obtener información rigurosa con la que establecer de forma precisa la *causalidad* de los cambios debidos exclusivamente a la intervención de cooperación para el desarrollo. Otra opción es acudir a diseños no experimentales (y cuasi-experimentales) en los que la asignación de pertenecer al grupo de tratamiento o control (en este caso suele denominarse grupo de comparación) no se realiza de forma aleatoria, es decir los miembros no tienen exactamente la misma probabilidad de pertenecer a uno u otro grupo[3].

Resulta, sin embargo, sorprendente que, así como en la medicina, farmacéutica y, en general, las ciencias biomédicas, las investigaciones experimentales están ampliamente difundidas y hay estrictos protocolos de seguridad y controles de calidad antes de ofrecer un tratamiento contra determinada enfermedad, en la cooperación para el desarrollo este tipo de diseños sean objeto de algunos recelos y apenas se estén comenzando a utilizar. La propia Metodología de Evaluación de la Cooperación Española I (MECE I o II en su caso, a partir de ahora) señala

---

información. Roche señala que "uno de los problemas más difíciles que deben tener en cuenta los procesos de evaluación de impacto: cómo resumir y presentar los resultados en uno o varios formatos útiles que no menoscaben la riqueza, diversidad y complejidad de la historia" (p.123).

[2] Por seguir con el ejemplo, puede suceder que la mejora del ingreso de un miembro de la familia tenga consecuencias de menor bienestar en otro miembro. Trabajar mas horas el varón supone mayor trabajo y tareas domésticas para la mujer. ¿Cómo operar con estos signos contrarios para determinar un efecto neto de la intervención? Ese es el problema de la agregación.

[3] Estos métodos se presentan de forma breve en MECE I, pp.74-79.

recelos éticos al uso de grupos de control: *"la dificultad para trabajar con grupos de control en la cooperación para el desarrollo es muy alta. Esto se debe tanto a los problemas prácticos que plantea la constitución de los grupos como a cuestiones éticas: ¿hasta qué punto es correcto privar a un grupo de personas de los beneficios del desarrollo sólo para poder evaluar el impacto de un proyecto de manera "científica"?* (MAE 1998:76). Más adelante comentaremos la validez de este planteamiento y trataremos de profundizar en alguno de los "problemas prácticos" que en la Metodología no se abordan ni se explicitan.

El resto del artículo se organiza como sigue. En la siguiente sección se delimita bien el concepto de impacto, mejor de impactos, pues son de múltiple naturaleza. En la sección 3 se exponen los pasos a seguir para poder realizar una evaluación que permita un juicio valorativo riguroso de los impactos. En la cuarta sección se valoran las fortalezas y debilidades de los principales diseños existentes para poder medir los impactos: los experimentos randomizados (o aleatorios), los diseños no experimentales y el enfoque participativo que utiliza herramientas evaluativas muy sencillas y rápidas centradas en los intereses de los propios usuarios de la ayuda. La quinta sección concluye con recomendaciones para la cooperación española.

2. QUÉ ES EL IMPACTO DE UNA INTERVENCIÓN.

La principal fuente de respuesta a esta pregunta en el contexto de la cooperación para el desarrollo la encontramos en el *Glosario de los principales términos sobre evaluación y gestión basada en resultados* (CAD 2002). Allí encontramos la siguiente información para el término **impactos** (*impacts*, tanto en inglés como francés):

"Efectos de largo plazo positivos y negativos, primarios y secundarios, producidos directa o indirectamente por una intervención para el desarrollo, intencionalmente o no" (CAD 2002:24).

Por su parte, la MECE II define el impacto como:

"el análisis de todo posible efecto o consecuencia de una intervención a nivel local, regional o nacional" (p.12) y lo relaciona con la eficacia en el sentido de ser mucho más amplio que ella ya que:

i) no se limita a los efectos previstos (incluye los imprevistos)

ii) no se circunscribe sólo a los efectos deseados (incluye indeseados)

iii) no se reduce al estudio de la población beneficiada.

Además aborda en problema de la atribución en los siguientes términos: *"se trata de identificar efectos netos, es decir, de comprobar la relación de causalidad entre la*

intervención y el impacto una vez que se dejan de lado las consecuencias provocadas por otras acciones, ya sean intrínsecas a la población analizada o provocadas por una política ajena a la intervención que se evalúa".

Ambas definiciones comparten:

- un carácter holístico; la evaluación de impacto (o de impactos) quiere recoger todos los efectos habidos en un lugar donde se ha intervenido;
- un carácter inclusivo de las intervenciones; no se excluyen las unidades menores de intervención (proyectos) que pueden ser evaluados en el impacto, ni se centran en grandes intervenciones, entendiendo por "grandes" de alto presupuesto o largo periodo temporal de ejecución (programas o políticas);
- un carácter selectivo en su atribución: hay que asignar valor y mérito a la intervención aislada de otras intervenciones que pueden sesgar los juicios;
- un carácter extenso: se habla de impactos previstos e imprevistos; programados o no; intencionados o no; directos e indirectos.

Podríamos completar esta caracterización especificando que los impactos recogen la multidimensionalidad de la realidad humana, tanto individual como grupal. Por tanto se deben analizar impactos económicos (monetarios, de ingreso, de consumo, financieros, productivos, etc.) y extraeconómicos (educativos, sanitarios, institucionales, políticos, sociales, entre otros). Es claro que todo análisis social tiene límites por lo que no se podrá profundizar en todos los aspectos con la misma intensidad. A menudo, el dilema intensidad *vs.* extensión del análisis deberá ser fruto del consenso que exige la negociación de los términos de referencia de una evaluación por parte de todos los agentes implicados (o *stakeholders*). Pero aunque se prioricen unas dimensiones de la realidad sobre otras, el análisis del evaluador debe ser lo más holístico o integral posible. Es decir, la evaluación de impacto supera enfoques más "ceñidos" como la evaluación por objetivos o incluso la evaluación de resultados.

La *evaluación por objetivos* se realiza dentro del marco analítico programado para la intervención y del compromiso asumido en relación con la asignación de recursos para dicha intervención. De alguna manera, puede centrar tanto la mirada del evaluador sobre la realidad programada en los objetivos que éstos se convierten en unas "orejeras" que limitan el campo de visión (análisis) del evaluador. Por ello Scriven (1973) desarrolla y defiende una metodología de evaluación libre de objetivos[4]. Hasta la fecha, la mayoría de las evaluaciones

[4] Para profundizar esta diversidad metodológica, véase Stufflebeam y Shinkfield (1993) o Larrú (2000) por ceñirse sólo a la literatura disponible en castellano.

que se realizan en España siguen este enfoque. A partir de la planificación realizada bajo el Enfoque del Marco Lógico en el que se definen objetivos general y específico, se esperan obtener productos (outputs) a partir de recursos previstos a financiar por la Cooperación. Es decir, en la actualidad tenemos acceso a información que nos permite saber si el dinero colocado en proyectos y programas ha sido invertido en lo que fue previsto. Pero esto nada nos dice sobre si las condiciones de vida concretas de los beneficiarios siguen igual, han mejorado o han empeorado. Este es el objetivo de la *evaluación orientada a resultados*. Ya no se centra la atención en el destino de los insumos, sino que se quiere conocer para qué ha servido la ayuda prestada. En qué ha cambiado la vida de las personas receptoras de esa ayuda, con independencia de que ese cambio se deba exclusivamente a la intervención o a otros muchos factores: contexto, políticas locales, mejora de la coyuntura nacional o internacional, migraciones, etc. Sólo cuando realizamos una evaluación que permite aislar y medir los efectos debidos a la sola intervención evaluada, estamos en condiciones de afirmar que es una *evaluación de impacto.* Puede afirmarse que mientras que la evaluación por objetivos tiene un punto fuerte en el control de las operaciones y la rendición de cuentas, la evaluación de resultados aporta conocimientos útiles sobre la eficacia de los procesos que generan desarrollo (o no) y la evaluación de impacto es muy potente en cuanto a la medición y atribución de éxitos (o fracasos). Notemos que ninguna de las tres garantiza de forma inmediata la replicabilidad de las acciones en otros contextos (la validez externa de las intervenciones). Pero se puede decir que la evaluación de impacto es la que mejor permite tener una base más firme para la reproducción de acciones, ya que además de saber si es exitosa o no, permite discriminar entre alternativas de diverso coste, debido a que ya sabemos que el efecto medido es causado sólo (o con un grado conocido de probabilidad) por la intervención.

Hay que reconocer que la historia de la cooperación para el desarrollo dispone de muy pocas evaluaciones que, con rigor, puedan considerarse de impacto. Como presentaremos más adelante, es a partir de 2004 que están comenzando a aparecer más publicaciones que, bajo diseños experimentales, sí son auténticas evaluaciones de impacto. La escasez de evaluaciones de impacto tiene causas técnicas, económicas y políticas.

La principal explicación *técnica* consiste en que, para poder valorar el impacto tal y como ha sido definido, es necesario contar con una información que permita calcular el *contrafactual* de la intervención. Esto es, cuál sería la situación de la población meta de no haberse ejecutado la intervención. Aquí se encuentra la justificación para crear los grupos de control (en los diseños experimentales) o de comparación (en los diseños no experimentales). El

grupo de control es precisamente la fuente de información del contrafactual de la intervención. Sobre esa población, se ha obtenido *la misma* información que sobre la población meta y, al haber sido asignada aleatoriamente, permite inferir que las diferencias entre ambas situaciones se deben exclusivamente a la intervención[5]. Otro aspecto técnico de importancia es la bondad de contar con una línea de base y un sistema de información riguroso que permita la detección de los cambios antes y después de la intervención. La línea de base no es imprescindible si el diseño de la evaluación se realiza bajo condiciones experimentales de aleatorización. Además, la línea de base por sí sola, no garantiza resolver satisfactoriamente el problema de la atribución, pudiéndose sobredimensionar o subestimar resultados, incluso bajo técnicas estadísticas sofisticadas, como se verá más adelante.

Dentro de las *causas económicas* destacan dos. La primera es la condición de bien público del conocimiento generado por las evaluaciones sobre lo que funciona o no, bajo qué circunstancias y a qué coste. Dada la naturaleza de no rivalidad en su consumo y no apropiación de los bienes públicos, los incentivos económicos para ofrecer evaluaciones de forma individual no satisfacen la demanda social de los mismos. Es decir, los beneficios sociales del aprendizaje generado por las evaluaciones rigurosas son mayores que los costes de afrontar en solitario dichas evaluaciones. Una agencia de desarrollo puede actuar como *"free rider"*, es decir, esperar a que sean otras quienes afronten el coste de evaluaciones y luego obtener beneficios de sus recomendaciones, ya que lo generado como conocimiento una vez no tiene por qué volver a repetirse. Esta condición de bien público es la que justifica la existencia de redes de evaluación como la de la OCDE (*The DAC Network on Development Evaluation)* o la propuesta para formar un Comité Internacional de Evaluaciones de Impacto, impulsada por el *Center for Global Development* (Savedoff, Levine y Birdsall 2006).

Otra razón económica importante que explica las pocas evaluaciones de impacto disponibles hasta la fecha, es la dificultad para encontrar financiación suficiente y adecuada en el momento oportuno. Por ejemplo, al iniciarse un programa social en un país en desarrollo y no conocerse de antemano los resultados que se obtendrán, el programa se ejecuta en una escala de fase inicial o piloto. En esta situación, es máxima la conveniencia de evaluar los cambios reales que produce dicho programa piloto antes de tomar la decisión de si se amplía a más zonas o incluso se generaliza a todo el territorio nacional. Es muy posible que existan varios

[5] Es necesario que la información recogida sobre ambas poblaciones o grupos tenga en cuenta algunas características que afronten problemas que se verán más adelante como son la atrición o contaminación entre los grupos de tratamiento y de control.

agentes involucrados en dicho programa pero que no quieran o no puedan aportar financiación para la evaluación. Si existieran fondos comunes para afrontar estas evaluaciones (por ejemplo dentro de un Comité Internacional de Evaluaciones de Impacto) ésta falta de financiación en un momento dado podría mitigarse o subsanarse.

Entre las *causas políticas* también pueden destacarse dos. En primer lugar, la falta de incentivos políticos. Es muy posible y probable que los gobiernos donantes no sientan presión ni necesidad por rendir cuentas de los *resultados* de los fondos de ayuda oficial al desarrollo (AOD), ya que el público (los pagadores de impuestos) y los medios de comunicación únicamente atienden al *insumo* de la cantidad comprometida de AOD (sobre todo en términos relativos de Renta Nacional Bruta y su cercanía o no al 0'7%)[6]. Además, los gestores y planificadores de la ayuda tienen mas discrecionalidad en sus asignaciones de fondos si el conocimiento sobre lo que realmente da resultado o no es bajo, ya que motivaciones distintas a la eficiencia en la reducción de la pobreza, tienen entonces mayores posibilidades de ser puestas en práctica.

En segundo lugar, la difusión y publicación de los resultados de las evaluaciones sufre de cierto sesgo hacia las experiencias positivas. Es decir, las revistas especializadas y las publicaciones académicas tienden a difundir experiencias exitosas mientras que los fracasos y sus causas quedan más ocultos. Salvo excepciones (y curiosamente más frecuentes en el campo de la ayuda humanitaria como por ejemplo el papel de la ayuda internacional y algunas organizaciones en el desastre de Ruanda y Burundi de 1994[7]), es difícil encontrar libros o artículos en las revistas internacionales de desarrollo que documenten fracasos y errores en la ejecución de intervenciones de cooperación para el desarrollo. Puede que la presión política ejercida por los diferentes agentes involucrados "depure" los resultados de los informes de evaluación o se vete su acceso al público.

De esta situación de déficit de evaluaciones de impacto que puedan ofrecer aprendizajes sobre los efectos realmente imputables a las intervenciones de cooperación para el desarrollo, se han hecho eco algunas publicaciones recientes. Entre ellas destacamos dos. En primer lugar, la iniciativa del Banco Mundial en 2004 a través de la Conferencia de Shanghai sobre reducción

---

[6] Sobre esta misma idea véase la provocadora obra de Easterly (2006:181-182).

[7] Véanse por ejemplo Storey (1997) y Uvin (1998). En Buchanan-Smith (2003) se comenta cómo la alta calidad del Estudio 3 del *Joint Evaluation of Emergency Assistance to Rwanda* ejerció una decisiva influencia en la creación del proyecto *Sphere* en 1996 para garantizar intervenciones de ayuda humanitaria que evitaran los errores cometidos.

de la pobreza (Moreno-Dobson 2005), que de 106 informes detectados como posibles oferentes de buenas prácticas, sólo 16 han podido ser clasificados como evaluaciones de impacto. En segundo lugar, la investigación llevada a cabo por el *Center for Global Development* de Washington que a partir de la detección de la escasez de evaluaciones de impacto rigurosas, ha publicado una propuesta para la formación de un "Comité Mundial de Evaluaciones de Impacto" en donde se pueda llevar a cabo el servicio de bien público que supone este tipo de evaluaciones (véase Savedoff, Levine y Birdsall 2006). En este informe se detalla cómo de 127 evaluaciones revisadas por la Organización Internacional del Trabajo, sólo dos contenían un diseño cuyas conclusiones fueran 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[8].

En resumen, realizar evaluaciones de impacto no es tarea sencilla, pero hay consenso creciente en su necesidad, su utilidad y su eficiencia frente a una ingente cantidad de evaluaciones (y de dinero) que no logran ofrecer la información realmente solvente para afrontar preguntas tan decisivas como si la ayuda al desarrollo logra realmente sacar de la pobreza a sus "beneficiarios", o –como planteamos al comienzo del artículo- qué es lo que realmente funciona, lo que no y a qué coste.

3. LA OBTENCIÓN DE LA INFORMACIÓN.

Una vez aclarado el concepto de impacto, en esta sección se presenta la condición necesaria para acometer cualquier tipo de evaluación, pero especialmente la de impacto: el levantamiento de la información pertinente y eficiente.

La clave de todo ejercicio de evaluación es la obtención de información. Si la información es de calidad, la evaluación será sólo la última etapa del denominado sistema de seguimiento (o monitoreo) y evaluación. Como es sabido, la información puede clasificarse en dos tipos: de

[8] 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 y que el 14% de los categorizados como de no impacto, eran realmente evaluaciones de impacto. Este dato puede reforzar la necesidad de este artículo de aclarar qué se entiende por impacto.

naturaleza cuantitativa o cualitativa. La primera mide y la segunda explica el significado de la medida.

Por una parte, la *información cualitativa*, será esencial escuchar cómo expresan los beneficiarios los cambios habidos en sus vidas durante y tras la intervención. Los testimonios suelen ser muy expresivos de la existencia de impactos positivos o negativos. Será interesante distinguir narraciones referidas a los productos del programa (actividades previstas realmente ejecutadas), de los relatos referidos a los efectos y los impactos (efectos sostenidos en el tiempo). Unas buenas entrevistas (a ser posible grabadas), unos grupos de discusión o muchas de las herramientas participativas pueden ser un material muy rico para el evaluador. Incluso puede pensarse en el uso de software específico para el tratamiento de textos narrativos. Por ejemplo, en un informe de evaluación se relataba cómo las mujeres que habían dejado de ser analfabetas narraban cómo les había cambiado la vida (los resultados de las capacitaciones realizadas): *"no se imagina lo que nos alegra poder leer las etiquetas de los alimentos, leer el número del autobús cuando nos desplazamos al zoco y, sobre todo, poder ayudar o saber lo que nuestros hijos hacen en el colegio"*. Sin este testimonio de resultado, la información que nos hubiera dado un típico informe final sería que el programa ha logrado hacer que 109 mujeres dejen de ser analfabetas, pero poco más.

Por otra parte, la *información cuantitativa*, si bien suele existir en la mayoría de los diseños previos o formulaciones de las intervenciones, la realidad es que, con frecuencia, su calidad deja mucho que desear. Los indicadores pueden ser incompletos o ineficientes. A menudo sólo se diseñan sistemas de información limitados a recoger el cumplimiento de las actividades programadas, pero no los resultados, ni los impactos. Para ello, conviene comprobar que los indicadores cumplen las características SPICED (subjetivo[9], participativo, interpretado, verificado, empoderador, diverso) o SMART (específico, inequívoco, sensible, pertinente y de duración limitada)[10].

A modo ilustrativo, consideramos el siguiente caso de formulación de un programa de desarrollo. El Objetivo específico es una mejora de los ingresos de la población meta. Para él se formula el siguiente doble Indicador Objetivamente Verificable (IOV):

---

[9] En el sentido de que, por ejemplo, los entrevistados clave o participantes directos en la intervención de desarrollo, tengan valoraciones únicas que, lejos de ser anecdóticas para el evaluador, se pueden convertir en cruciales a la hora de juzgar el mérito y valor de la intervención.

[10] Una descripción mas extensa de estas propiedades puede consultarse en Roche (2004:70-72).

- *"2 actividades productivas locales se articulan para incrementar el nivel de ingresos de las comunidades rurales"*
- *"1 actividad económica (turismo) se reconoce como generadora de ingresos"*.

Las Fuentes de Verificación (FV) serán:

- *"Estudios de la delegación de Agricultura sobre la producción y comercialización (almendras)"*
- *"Informes de la Delegación de Turismo, memorias de entrevistas con Delegados étnicos"*
- *"Informes de evaluación del programa"*

Los riesgos previstos son:

- *"Se incrementa por encima de lo habitual la tasa de emigración de la población de las comunidades rurales de la comunidad XX"*
- *"Se producen tensiones políticas en el país que impiden el normal movimiento de personas entre regiones limítrofes"*.

Según éstos parámetros, parece ser que los ingresos de la comunidad (¿toda?) dependen de la articulación de dos actividades productivas y del turismo. En principio son dos medios que potencialmente pueden incrementar los ingresos, pero existe un alto grado de desinformación e incertidumbre no reflejada en los riesgos: ¿cuál es el nivel de ingresos de partida para poder compararlos tras el programa? ¿Cuál es el alcance del programa? ¿Los ingresos que deben incrementarse son los de los beneficiarios o los de toda la comunidad? El diseño de evaluación de impacto sería distinto en cada caso. En el primero, el nivel de agregación de la información serían encuestas de ingresos familiares a beneficiarios y no beneficiarios (grupo de control); en el caso de toda la comunidad, habría que diseñar una comunidad de control lo más exacta posible a la de tratamiento.

Parece más realista suponer que los ingresos son función de un mayor número de variables exógenas al programa: empleo en la zona, infraestructuras de comercialización, demanda interna, macroeconomía del país, exportaciones, etc.

Incluso el turismo es función de más variables que la emigración o movilidad interna. La demanda interna será función –al menos- de la renta de los ciudadanos, su cultura vacacional, la competencia con otros destinos alternativos competidores... y la demanda externa de factores como la seguridad, el tipo de cambio, precios relativos de destinos competidores, infraestructuras, profundidad de las ofertas de tour operadores, etc.

Más allá de esta sobreambición común en muchos objetivos de proyectos y programas de desarrollo, es claro que no se ha diseñado un sistema de información que permita juzgar el resultado concreto sobre la pobreza de ingresos. Es más, este tipo de información es la que

debe desprenderse de la Línea de Base, sobre todo cuando ya se financia un programa aprobado y no se pide de forma previa. Encuestas de ingresos de hogares a los beneficiarios y un grupo de control es posible y con un coste razonable para conocer el impacto. La recomendación que podría hacerse en este caso es simplificar los objetivos y diseñar sistemas de información (indicadores) que realmente superen la realización de actividades para orientarse a resultados. En su caso, la institución encargada de la valoración ex –ante de los programas deberá tener esto muy en cuenta. Una vez aprobado el programa, deberá volverse a diseñar los IOV de forma rigurosa, especificando en las FV quién debe recoger esa información, cada cuánto tiempo y qué presupuesto tiene asignado. Los defectos o carencias de los IOV deberían apreciarse por el donante y los ejecutores del programa, ya durante el seguimiento, no debiendo esperar a que el informe de evaluación final excuse el impacto o resultados por carecer de la información necesaria.

Para ilustrarlo mejor, consideremos ahora este otro ejemplo de IOV:

"de las 60 mujeres beneficiarias de acciones en turismo rural en la comunidad XXX cuyos ingresos de partida son YYY unidades monetarias locales (línea de base que contiene sujetos, lugar y cantidades), *un 50%* (30 mujeres) (impacto cuantitativo) *encuentran empleo en turismo rural* (impacto cualitativo) *y obtienen unos ingresos adicionales de ZZZ unidades monetarias locales* (impacto cuantitativo) *tras un periodo de cinco años* (tiempo)".

Como se aprecia se considera un IOV bien formulado aquél que contiene descrita la situación de partida en cantidad, espacio y beneficiarios, una medida de impacto cuantitativo y/o cualitativo esperado, en un plazo de tiempo esperado concreto.

Por su parte, la FV bien formulada podría ser: *una* encuesta *semestral durante cinco años a las 60 mujeres beneficiarias sobre sus ingresos; si culturalmente es posible,* nóminas *de ingresos;* contratos *de nuevo empleo; encuesta o registro de ingresos de 60 mujeres no beneficiarias lo más parecidas a las beneficiarias (edad, lugar de residencia, formación, etc.).* Recogidas por *el director del programa con ayuda de estudiantes en turismo.* Presupuesto: *XXX unidades monetarias*.

Nótese que además de la descripción precisa de actividades a realizar, se ha concretado quién va a ser el responsable de su recogida y existen medios económicos para realizarlo[11].

Una información de suficiente calidad y un diseño de intervención que presente con claridad y rigor una hipotética cadena causal son dos premisas necesarias -y no suficientes- para afrontar

[11] Para profundizar de forma práctica en la formulación de IOV véase Gómez Galán y Cámara (2003:33-36).

la medición del impacto. Es necesario recalcar que sin datos de calidad, cualquier evaluación va a perder mucha credibilidad y utilidad.

Tanto en los diseños experimentales y no experimentales, los datos recogidos en el grupo de tratamiento y de control (o comparación), permitirán trabajar con una base de datos que debe irse limpiando a medida que se actualiza. Todo el tiempo y esfuerzo que se invierta en este momento, será inversión productiva a la hora de generar estimaciones de alta calidad, rigor y credibilidad. A mi juicio, las dos debilidades más frecuentes en los informes de evaluación son unas conclusiones débilmente argumentadas (quizá veraces, pero no lógicamente válidas) y la ausencia de las pruebas (contraste o verificación) "objetivas" que proporcionan los datos. Sin una base de datos extremadamente fiable, el análisis a partir de ella estará sesgado, por más sofisticadas técnicas estimativas que usemos.

El primer paso para elaborar una eficiente base de datos es determinar la cantidad de datos a recoger, es decir, determinar el tamaño, tipo y poder muestral.

Los tipos muestrales se encuentran fácilmente tratados en la literatura estadística (aleatorios, estratificados, etc.)[12]. Pero una cuestión esencial será conocer el tamaño muestral necesario para poder afirmar con rigor conclusiones precisas. Es lo que se conoce como el *poder muestral* o la probabilidad de detectar un efecto en la muestra seleccionada, cuando realmente se ha producido en la intervención de desarrollo que se evalúa. El poder muestral será función de factores como el nivel de agrupamientos (clusterización) a realizar en la investigación evaluativa, la disponibilidad de una línea de base previa o no, disponibilidad de variables de control y estratificación, así como del tipo de hipótesis a testar. Desde el punto de vista práctico, lo común es usar un software que permite obtener el número de clusters necesarios para niveles de confianza determinados (por ejemplo el *optimal design* creado por la Universidad de Michigan[13]). Dicho software requiere como insumos para sus cálculos que el evaluador determine valores para los siguientes parámetros:

- nivel de confianza deseado (α)
- número de clusters con los que se va a trabajar (n)
- efecto estandarizado (efecto esperado dividido entre la desviación típica) (δ)[14]

---

[12] Al respecto, y con ejemplos de evaluación de intervenciones de desarrollo, puede consultarse Casley y Kumar (1990).

[13] http://www.optimaldesign.org/software.html Véase Raudenbush, S. et al. (2004).

[14] Los valores mas usuales son 0,2 para apreciar un efecto pequeño, 0,4 para un efecto mediano y 0,5 para grande.

- correlación entre las unidades del cluster (ρ).

Una vez que tenemos la base de datos depurada a la máxima credibilidad y ajustada al tamaño fijado para el poder muestral requerido, puede procederse a la estimación de los impactos. Para ello disponemos de varias alternativas metodológicas que se presentan a continuación.

4. CONTRASTE VALORATIVO DE LOS POSIBLES DISEÑOS PARA MEDIR LOS IMPACTOS.

En esta sección vamos a considerar tres paradigmas o enfoques evaluativos que difieren, entre otras características, en el tipo de información que priorizan. Los dos primeros son el enfoque y diseño experimental y el no experimental. Tradicionalmente, las ciencias han utilizado diseños experimentales (ciencias biomédicas y materiales) y no experimentales (ciencias sociales) para construir su conocimiento. Todas trabajan con herramientas cuantitativas y cualitativas, con lo que es erróneo identificar el diseño experimental sólo con ciencias materiales (la física fue el más admirado) y con la información cuantitativa, mientras que las ciencias sociales deben limitarse al diseño no experimental y las herramientas cualitativas. El tercer enfoque es el que aquí denominamos participativo. Su interés por la información es utilitario, pues lo que pretende es que sean los propios beneficiarios de una intervención, su realidad concreta, contextual, diversa y subjetiva, los que logren realizar el ejercicio evaluativo y así lograr un mayor empoderamiento. Lo que se pretende en esta sección es valorar las ventajas e inconvenientes de cada uno de ellos, resaltando las diferencias en el tipo de información y conocimiento que reportan. Para una mejor comprensión de estas diferencias se utilizarán ejemplos de evaluaciones del sector educativo.

4.1 Evaluaciones de impacto bajo diseños experimentales.

Este tipo de evaluaciones, también denominadas aleatorias, randomizadas o experimentos de campo (*field experiments)* presentan la enorme ventaja de que las diferencias entre los grupos de tratamiento y de control son atribuibles únicamente a la intervención y no a otros factores o sesgos. Aleatorizar consiste en verificar que todos los miembros participantes en la intervención tienen la misma probabilidad de ser elegidos sea como tratamiento o como control. Esto puede realizarse de varias maneras, desde lanzar una moneda al aire y según el resultado para cada participante asignarle al grupo “cara” el tratamiento y “cruz” el control, hasta listar alfabéticamente a todos los participantes y asignar la aleatorización con la ayuda de un generador de números aleatorios (la mayoría de las hojas de cálculo informáticas ya disponen de esta función). Conviene resaltar que no es lo mismo una *asignación* aleatoria de

los participantes que una *muestra* aleatoria. Mientras la primera afecta a la validez interna de la evaluación, la segunda afecta a su validez externa.

Gustafson (2003) propone denominar al lanzamiento de la moneda o la asignación por números aleatorios como *aleatorización completa (full randomization)* y advierte que existe otra posibilidad: una aleatorización restringida *(restricted randomization).* Esta consiste en generar aleatorización pero definiendo previamente el número de componentes que va a tener cada grupo, por ejemplo 10 miembros de grupo de tratamiento y 10 para el de control. Esto evita que las dimensiones entre ambos grupos puedan ser muy diversas en tamaño si han sido generadas por aleatorización pura o completa. Para el caso de intervenciones de desarrollo es muy interesante, pues mediante la aleatorización restringida aseguramos que el número de beneficiarios y "perjudicados" (en el sentido de no recibir el beneficio previsto por la intervención al ser grupo de control), es exactamente el mismo. Gustafson prueba que ambos tipos tienen la misma potencia estadística. Además, la restringida permite controlar mejor los casos de *atrición* o desgaste (miembros del grupo de tratamiento que renuncien a seguir participando una vez iniciado el programa), los de *contaminación* (miembros del grupo de control que pasan a participar) y los *efectos derrame (spillovers)* por los beneficios inducidos del tratamiento.

Estos sucesos, conocidos como sesgos en la selección muestral, permiten diferenciar dos tipos de impactos. Uno es el conocido como impacto de *tratamiento intencional (intention to treat effect).* Es decir, estimamos el impacto sobre el grupo de tratamiento completo, incluyendo atrición y spillovers. El otro, denominado tratamiento sobre los sujetos realmente tratados *(treatment on the treated),* considera los casos de atrición y spillovers. Kremer y Miguel (2004) ofrecen una interesante investigación de cómo los impactos fueron muy diferentes en un programa educativo con tratamiento antiparasitador con alumnos de colegios en Kenia. Gracias a las externalidades producidas por la falta de contagio entre los alumnos tratados (*treatment on the treated impact*), los beneficios en la salud y asistencia a clase del total de la población escolar se estimaron un 61,1% superiores (impacto del tratamiento intencional) a los detectados únicamente en los alumnos que conformaban el grupo de tratamiento inicial. En otras palabras, si sólo medimos los impactos sobre el grupo de tratamiento inicial y no tenemos en cuenta la atrición, contaminación y externalidades del programa, podemos estar obteniendo subestimaciones o sobreestimaciones importantes.

La objeción sobre la moralidad de los diseños experimentales.

La evaluación por experimentación en intervenciones de cooperación para el desarrollo ha recibido críticas como las ya señaladas en el epígrafe 1 de este trabajo (reparos morales, sobre todo). Estas cautelas son ahora compartidas en el campo de la medicina o farmacología, donde han sido utilizadas desde hace años y ahora se buscan desarrollar protocolos que pauten la actividad, pero pocos la rechazan frontalmente y niegan su utilidad en el avance del conocimiento sanitario. Autores como Duflo (2005) o Banerjee (2006) han defendido las evaluaciones aleatorias como el mejor camino para aprender y construir conocimiento cierto sobre el impacto y efectividad de las intervenciones de desarrollo. La evaluación aleatoria invita a la creatividad de teorías y desafía muchas de las "sospechas" convencionales, sobre todo en un campo tan debatido como la eficacia de la ayuda (véase Easterly 2006).

El reparo moral a este tipo de evaluaciones puede expresarse como la duda de si es ético negar un tratamiento (sobre todo ante situaciones de pobreza) a un grupo de personas y "utilizarlas" como conejillos de indias mediante el grupo de control.

Frente a esta objeción, lo primero que debe advertirse es que no siempre es posible, ni conveniente, realizar evaluaciones experimentales. Cuando el tratamiento se realiza sobre el universo poblacional, no es posible crear un grupo de control (por ejemplo, políticas de alcance completo en una ciudad, políticas económicas como el tipo de cambio óptimo para un país o la política de pensiones).

Un segundo reparo a la objeción moral es que en los diseños experimentales, los que forman el grupo de control no son reducidos a "conejillos de indias", sino que se les trata o compensa de otra forma (se les presta un servicio alternativo y no correlacionado con el que se está evaluando, el placebo en medicina), o la situación entre los componentes de tratamiento y control se invierte pasado un periodo de tiempo. Por ejemplo, en la evaluación conducida por Banerjee et al. (2007), hubo una rotación en las clases que obtenían el tratamiento (una persona de refuerzo que atendía a los alumnos "tratados" durante dos horas denominada Balsakhi –literalmente amigo de los niños- en la India). El programa se implantó en tercer y cuarto grado a lo largo de dos años y cada clase se dividió en dos grupos. El diseño evaluativo se muestra en la Figura 1.

Figura 1. Diseño experimental de un programa de mejora educativa en la India.

	Año 1		Año 2		Año 3	
Ciudad 1	*3er curso*	*4º curso*	*3er curso*	*4º curso*	*3er curso*	*4º curso*
Grupo A	T	C	C	T	C	C
Grupo B	C	T	T	C	C	C

Ciudad 2						
Grupo C	T	C	C	T	C	C
Grupo D	C	C	T	C	C	C

Fuente: Elaboración propia a partir de Banerjee et al (2007). T=tratamiento y C=control.

De esta forma, casi todos los alumnos necesitados de refuerzo pudieron beneficiarse del Balsakhi en algún momento del cronograma y la información se siguió recogiendo durante el año siguiente a la finalización de actividades para apreciar los efectos de sostenibilidad. Esta evaluación produjo conocimiento preciso sobre cómo mejorar la calidad educativa. El grupo de alumnos que recibieron el tratamiento, mejoró sus calificaciones en los test estándar en un 0,14 de desviación típica frente a la situación inicial el primer año, y en un 0,28 de desviación típica el segundo. Los alumnos más atrasados fueron los que más mejoraron. El tercio peor de alumnos mejoró en 0,4 desviación estándar el segundo año. En conjunto, y estimando bajo una variable instrumental para controlar el desgaste y los efectos indirectos, el programa proporcionó una mejora media de 0,6 desviaciones estándar. Los autores comparan la eficiencia del bajo coste de esta iniciativa (los Balsakhi recibían 10-15 dólares al mes) frente a otras iniciativas educativas como el programa STAR de Tennessee que, siendo mucho más caro, intentó mejorar las notas de alumnos reduciendo el tamaño de las clases de 22 a 15 alumnos y sólo lo consiguió en un 0,2 de desviación típica (Krueger & Whitmore 2001), o las iniciativas de mejorar la calidad de la enseñanza mediante *flip charts* en Kenia evaluada por Glewwe et al. (2004) que retrospectivamente mostraban un logro de 0,2 desviación típica de mejora en las notas, pero la revisión aleatoria prospectiva no encontró mejora alguna. Otra evaluación midió la posibilidad de incentivar a los profesores mediante premios si sus alumnos mejoraban la calificación en los test. El sorprendente resultado, que se reveló no sostenible en el tiempo, fue que los profesores sólo acentuaban la preparación de los alumnos para los test, pero no hacían un mayor esfuerzo docente que produjera cambios sostenibles (Glewwe et al. 2003). El informe de esta evaluación es muy interesante pues compara entre sí tres buenas ideas para mejorar el rendimiento escolar en Kenia: mejorar la dieta de los alumnos reforzando el desayuno que se ofrece en los colegios, regalar los libros de texto y el uniforme para que estudien mejor y/o no falten a clase, o tratar masivamente a los alumnos contra las enfermedades parasitarias. La primera acción logró aumentar la asistencia a clase en un 30% costando 36$ por año y alumno. La compra de uniformes y libros de texto, costaba 99$ y logró mejorar un 15% la asistencia y únicamente las notas de los mejores alumnos (lo que se explica porque los libros estaban editados en inglés y sólo la minoría que dominaba

esta lengua pudo aprovecharlos, coincidiendo en ser miembros de las castas superioires que ya obtenían buenos resultados). La desparasitación logró aumentar la asistencia a clase un 25% costando únicamente 3,5$ por alumno.

La objeción moral al diseño experimental es interesante porque también remite al modo en que se seleccionan los beneficiarios en las intervenciones de desarrollo. ¿Cómo se decide en qué comunidad se interviene y en cuál no, con qué personas se trabaja –beneficiarios- y cuál no? Si le preguntáramos a la práctica habitual, obtendríamos respuestas que conducen a sesgos importantes en la selección y quizá no siempre justificados. A menudo, las organizaciones y agencias donantes aseveran que sus intervenciones se dirigen (o se trata) a los *más* pobres de cada comunidad, pero casi nunca se prueba (¿cómo se ha medido o apreciado que son los más pobres? ¿Se ha diagnosticado alcanzando al universo poblacional de la comunidad o del país?). Lo que, en mi opinión y experiencia suele hacerse, es un esfuerzo de diseminación de la intervención que se va a realizar en una comunidad previamente elegida de forma no aleatoria y se anima a la gente a que se apunte voluntariamente a intervenir si cumple ciertos criterios fijados por la organización. Pero no hay garantía de que dichos criterios sean siempre cumplidos con pulcritud ética y bien pueden ser considerados subjetivos. En el fondo, la mayoría de las veces el tratamiento lo recibe quien ha tenido la suerte de tener acceso a esa información (que nunca será completa), sea seleccionado por la organización (bajo sus criterios "objetivos" pero establecidos de forma subjetiva en el diseño de la acción, a menudo no realizado de forma participativa con *todos* los miembros de una comunidad). Frente a esta práctica, la aleatorización no fija ningún criterio selectivo a priori, sino que todos los potenciales beneficiarios son repartidos de forma aleatoria entre tratamiento y control, con la idéntica probabilidad de ser seleccionado en uno u otro grupo. De esta forma, la aleatorización, como mecanismo para una implementación que no discrimina entre segmentos de la población, resulta más transparente y justa que modalidades de implementación donde, ya sea un burócrata, un técnico de una ONG o "un experto de reconocido prestigio", decide quien recibirá el tratamiento de la intervención. La aleatorización, por tanto, genera como subproducto condiciones altamente apropiadas para realizar una evaluación rigurosa.

Otra respuesta a la objeción ética, consiste en el realismo de las restricciones presupuestarias y de recursos del programa. De ahí el que a menudo se inicien programas piloto que, una vez evaluados, puedan extenderse a otros lugares. Es el caso del programa de transferencias en efectivo llevado a cabo en México (Progresa en su versión primigenia y actualmente denominado Oportunidades) que tras los exitosos resultados evaluados de forma aleatoria,

permitió extender los recursos a muchos más estados mexicanos, sobrevivir a varios cambios de dirigentes políticos y hoy se ha adaptado a países como Brasil ("Bolsa Escola"), Colombia, Perú o Sudáfrica[15]. Pero, ¿hubiera sido ético (bueno) extender este programa sin saber exactamente si funcionaba o no? ¿Cuál es el coste de oportunidad de la ignorancia de los impactos? ¿Es ético seguir dando dinero y ejecutando proyectos que no sabemos si realmente reducen la pobreza? Por ejemplo, soy testigo de muchos informes sobre proyectos y programas de capacitaciones laborales o de microempresas, que no reportan ningún resultado concreto en su objetivo específico de aumentar los ingresos de los beneficiarios. Por el contrario, se acumula evidencia de los grandes problemas que existen cuando se quiere comercializar los productos que se han diversificado por los beneficiarios, pero para los que o no hay mercado, o no hay infraestructuras adecuadas, o son mercados saturados por una oferta excesiva y no diferenciada a la demanda local.

En resumen, si bien no debe mantenerse una postura maximalista que convierta al diseño experimental en la panacea del desarrollo y se convierta en "la única" forma de evaluar, ya no puede mantenerse que no sea posible por el simple hecho de que en ciencias sociales la libertad humana lo imposibilite (no se experimenta en laboratorios que aíslan la variable dependiente de posibles influencias externas, es decir las demás variables endógenas), o que "en todos los casos" sea desaconsejable por cuestiones éticas. Más bien parece que lo aconsejable es intentar realizar evaluaciones experimentales de impacto *siempre que se pueda* y acudir a los diseños alternativos cuando la situación o las circunstancias contextuales impidan acometer el enfoque aleatorio.

4.2 Evaluaciones de impacto bajo diseños no experimentales: ¿cuánto nos podemos equivocar?

Por lo que respecta a los métodos no experimentales, los más usados en las evaluaciones son la estimación bajo diferencias en diferencias (véase Duflo & Kremer 2005 para una crítica de sus debilidades frente a la aleatorización); el pareo (o *propensity score matching)*; el uso de variables instrumentales, la discontinuidad en la regresión o el truncamiento en series temporales. Manuales de desarrollo como el Baker (2000) o el trabajo de Ravallion (2005) hacen una buena descripción de sus potencialidades y limitaciones en las que, por extensión, aquí no nos podemos detener.

---

[15] Para las evaluaciones de dicho programa consultar Skoufias (2005), Djebbari & Smith (2005), Schultz (2004), Hoddinot & Skoufias (2004) o véase http://www.ifpri.org/spanish/pubs/spubs.htm#progresa.

Sí queremos resaltar que lo mejor es utilizar el diseño experimental en la medida de lo posible, ya que las revisiones realizadas de evaluaciones bajo distintos diseños no experimentales frente a los experimentales, arrojan impactos tremendamente diferentes. Para hacernos una idea de la importancia de los sesgos y errores que introducen los diseños no experimentales, sintetizamos los trabajos de LaLonde (1986) y Arceneaux et al. (2006).

Lalonde analizó los resultados de un programa de inserción laboral e incremento de ingresos para personas excluidas del mercado laboral. Los seleccionados aleatoriamente en el grupo de tratamiento, recibían formación y apoyo en pequeños grupos y se les conseguía un empleo de entre 9 y 18 meses. Los miembros del grupo de control únicamente recibían la formación, pero no el empleo temporal. La comparación que hizo Lalonde entre los resultados del diseño experimental y dos técnicas econométricas alternativas se resumen en la siguiente tabla:

Tabla 1. Incrementos de ingreso (dólares de 1982) ante varios diseños estimativos.

	Grupo de control	**Varones**	**Mujeres**
Diseño Experimental		798 (472)	861 (306)
No Experimental en un paso	Tipo 1	-1.228 (869)	2.097 (491)
	Tipo 2	-805 (484)	1.041 (505)
No experimental en dos pasos	Tipo 1	-1.333 (820)	1.129 (385)
	Tipo 2	-22 (584)	1.102 (323)

Fuente: seleccionados a partir de LaLonde (1986). Los resultados del diseño experimental están ajustados por las exógenas edad, edad al cuadrado, años de escolarización, abandono escolar antes del obligatorio y raza. El grupo de control tipo 1 son todos los varones/mujeres que participaron entre 1975-78, menores de 55 años y no jubilados. El control tipo 2, incluye a los mayores de 55 años. Para la estimación de las mujeres se usaron 8 tipos de grupo de control y para los varones, 6. Entre paréntesis, los errores estándar.

La principal conclusión de este ejercicio de sensibilidad es que los diseños no experimentales, ofrecen resultados sistemáticamente más positivos y altos en las mujeres y más negativos y bajos en los varones, que los experimentales. Los diseños en dos etapas, tienen menos errores de especificación que los de una sola, pero los experimentales no poseen este tipo de error al ser independientes de la especificación del modelo. ¿Cómo calificar a un programa que, en función del control y especificación elegidos, reporte aumento de ingresos de más de 2.000 dólares o de disminución de 1.300?

Arceneaux y otros (2006) han replicado los resultados de una campaña a favor del voto en los Estados Unidos, utilizando varios estimadores. Las diferencias frente al diseño experimental quedan reflejadas en la Tabla 2.

Tabla 2. Impacto estimado en una campaña de animación al voto.

Método de estimación	Impacto en %
Diferencia simple	10,8
Regresión múltiple	6,1
Regresión múltiple con datos de panel	4,5
Pareo (matching)	2,8
Experimento aleatorizado	0,4

Fuente: Arceneaux et al. (2006).

Como se aprecia en la tabla, todos los estimadores no experimentales sobrevaloraron el resultado de la campaña frente al uso del diseño experimental. Es una muestra clara de cuán equivocados podemos llegar a estar en los impactos de una intervención si, pudiendo realizar un diseño experimental, no optamos por él.

4.3 Un paradigma alternativo: el participativo.

Este paradigma pretende resaltar la conveniencia de que sea la realidad de los propios beneficiarios la que se ponga en el primer plano de los intereses de la evaluación (Chambers 1997) con el fin de que toda evaluación sea un ejercicio de empoderamiento de los propios usuarios de la ayuda (Fetterman 2001). Es cierto que el interés predominante de los diseños "clásicos", tanto experimental como no experimental, ha sido generar conocimiento para los promotores de la evaluación, normalmente los donantes (agencias, gobiernos u organismos internacionales de los países desarrollados).

Pero la utilidad de dichos estudios evaluativos para los receptores -los pobres- es, como poco muy limitada y como mucho, no pasa de ser tratados como simple fuente de información en la recogida de los datos o las entrevistas y demás herramientas cualitativas. No es corriente ver recomendaciones dirigidas a los beneficiarios en los informes de evaluación, ni suele existir una simple reunión de comprobación de que se han recogido bien sus conclusiones y juicios evaluativos.

Como puede deducirse, frente al enfoque experimental, el participativo no pone el énfasis en medir, aislar o atribuir causalidad, pero –según sus defensores- también es capaz de ofrecer impactos (Mayoux y Chambers 2005). El objetivo de este tipo de evaluaciones no es tanto medir con exactitud la causalidad de un programa, sino provocar que los participantes comprendan mejor su propia situación, promover el entendimiento mutuo llegando a conclusiones válidas para el grupo, mediante un análisis participativo, equitativo y fortalecedor de la propia comunidad, creando redes para futuras investigaciones. Para ello,

emplean herramientas propias, simples, rápidas, baratas y –sobre todo- centradas en la participación de los beneficiarios. Para conducir este tipo de evaluaciones, se requieren evaluadores muy entrenados en actitudes más que en conocimientos técnicos. Conductas y habilidades empáticas que promueven la dinámica grupal participativa. Como reconocen Mayoux y Chambers (2005), la utilidad de estas evaluaciones depende mucho de cómo se usen, por quién y de la voluntad política de considerar sus resultados. De forma sintética, los componentes principales del este paradigma se ofrecen en la Figura 2.

Figura 2. Síntesis del paradigma participativo de evaluación de impacto.

¿QUÉ ESTÁ PASANDO?
–¿Cuáles son los cambios prioritarios para los pobres?
–¿Cómo traducirlos en indicadores?
–¿Cómo comparar y agregar entre grupos y áreas tan diversas?
¿PARA QUIÉN EVALUAMOS?
–¿Qué diferencias de impacto hay entre los pobres?
–¿Cuáles son sus principales líneas y conflictos de intereses?
¿POR QUÉ HAN OCURRIDO ESTOS HECHOS?
–Superar la linealidad y comprender lo complejo del cambio para reducir la pobreza
–Comprender las interacciones estrategias-programas-contexto
¿QUÉ DEBERÍA HACERSE?
–¿Qué quiere hacer la gente? ¿Cómo evaluar los dilemas (trade-off) entre prioridades?
–¿Cómo negociar las diferencias y conflictos?
–Cuáles son las lecciones del pasado? Las oportunidades actuales? Los retos futuros?
–¿Quién necesita saber qué para tomar decisiones y actuar?

Fuente: Basado en Mayoux y Chambers (2005)

El principal reto al que se enfrentan estas evaluaciones es el mostrar que su conocimiento supera la utilidad de la participación local. Es decir, que también genera insumos válidos para la rendición de cuentas de los financiadores de los programas y que el conocimiento generado supera la simple subjetividad de un proceso opinático comunitario conducido por un "facilitador experto". Además, comparte con el resto de los enfoques las dudas acerca de la validez externa de sus conclusiones, pero sin alcanzar la causalidad y medida de la eficiencia que los diseños experimentales llegan a mostrar.

5. CONCLUSIONES Y RECOMENDACIONES PARA ESPAÑA.

¿Estaríamos en condiciones de responder con certezas probadas a un tomador de decisiones políticas que nos preguntara qué es lo que funciona y lo que no en la cooperación para el desarrollo? Probablemente todavía no. Pero se ha iniciado un interesante camino en la

comunidad científica en torno a la realización de evaluaciones de impacto que, a través de la aleatorización propia del diseño experimental, está empezando a ofrecer resultados concretos muy interesantes, convirtiendo la atribución, la causalidad y el contrafactual en algo posible en ciencias sociales. No es el único modo de hacer evaluaciones, ni son posibles en cualquier circunstancia. Lo que aquí se ha tratado de recalcar es que es la mejor forma de obtener conocimiento evaluativo, cuando sea posible hacerlo.

España aún no dispone de ningún ejercicio evaluativo de este calibre. Para superar esta ausencia, es probable que lo mejor sea acercarse de forma conjunta a quien ya está lo está haciendo, aprender haciendo (*learning-by-doing*) y cofinanciar evaluaciones de impacto internacionales que constituyen un bien público. Adherirse a iniciativas como la creación de un Centro Internacional de Evaluaciones de Impacto descrita en Levine y Savedoff (2005), es una cuestión de valoración y decisión política. Otra iniciativa de gran interés es la surgida dentro del Grupo Independiente de Evaluación del Banco Mundial, denominada *Development Impact Evaluation Initiative* que actualmente investiga sobre cinco áreas prioritarias: gestión educativa y participación comunitaria en los colegios; información y rendición de cuentas en educación; contratos e incentivos a profesores; transferencias corrientes condicionales a programas que incrementen los resultados en educación; y educación en programas para barrios marginales. Apoyar financieramente estas evaluaciones y aprender de sus resultados es otra decisión política que España puede plantearse.

Aunque las evaluaciones randomizadas sean las más potentes en medición y validez interna, los diseños no experimentales y el enfoque participativo son complementos valiosos para afrontar el problema de la validez externa o reproducción de resultados en otros ámbitos. Pero es muy importante que todas las evaluaciones vayan introduciendo cada vez más la dimensión cuantitativa. Sería interesante ver cómo en España no sólo se publican los informes evaluativos, sino que se ponen a disposición las bases de datos generadas a lo largo de los proyectos y programas, para que los investigadores puedan replicar resultados y aprender de las experiencias vencidas. Para que esto llegue a ser realidad, habrá que empezar por ser más exigentes en los sistemas de información construidos para el seguimiento y evaluación de los proyectos y programas. Sobre todo en la calidad de formulación de los indicadores y sus fuentes de verificación. Sería conveniente que aquellas propuestas que no alcancen un rigor mínimo en los indicadores de resultados esperados, no fueran financiadas. Además, España puede aprovechar la oportunidad para avanzar en evaluaciones de resultados que superen la rendición de cuentas de las actividades realizadas, y puedan recoger los cambios producidos en la vida de las personas y sus posibles factores influyentes, antes de acometer una

evaluación de impacto randomizada como la aquí descrita. Aquellos programas piloto y de cierto alcance de la AECI y la obligatoriedad de evaluación de los convenios de AECI con las ONGD, son momentos muy oportunos para elevar el nivel de las evaluaciones españolas y ofrecer lecciones de buenas y malas prácticas.

Es mucho lo que falta por saber en la eficacia de la ayuda al desarrollo. Lo que sí sabemos ya es que las intervenciones de desarrollo no pueden copiarse sin más, sino que deben adaptarse a cada contexto siendo ésta una variable esencial en toda investigación evaluativa. No debemos convertir ningún medio -por potente que sea- en una panacea que, a modo de barita mágica, creamos que va a solucionar la pobreza en todas partes. Incluidas las evaluaciones aleatorias. Lo ha dejado bien mostrado Easterly (2001) atacando las inversiones, la educación, el control demográfico, el ajuste estructural o la condenación de la deuda externa, no porque no sean importantes en el desarrollo, sino porque al considerarlas panaceas, se han rebelado estrategias insuficientes y fracasadas. Lo esencial es identificar las buenas preguntas *(enduring questions)*, usar métodos que produzcan respuestas científicas y evaluar con independencia para que los impactos sean creíbles, atribuibles y se puedan poner en marcha las recomendaciones emanadas de la evaluación. Publicar informes que ofrezcan lecciones de los errores para atenuar el sesgo positivo de las publicaciones únicamente exitosas que hoy existe es otro reto. Afrontarlo parece sencillo aunque sea arriesgado políticamente. Basta con publicar también los errores (tan frecuentes es un proceso tan complejo como el desarrollo humano) para que no se cumpla el viejo dicho de que el hombre es el único animal que se permite tropezar varias veces con la misma piedra (y no ser capaz de decirle a nadie que basta con apartarla del camino o dar un rodeo).

REFERENCIAS BIBLIOGRÁFICAS.

ARCENEAUX, K.; A. GEBER & D. GREEN (2006) "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment", *Political Analysis 14 (1), 37-62.*

BAKER, J. (2000) *Evaluación del impacto de los proyectos de desarrollo en la pobreza. Manual para profesionales.* Banco Mundial. Washington, D.C.

BANERJEE, A. (2006) "Making Aid Work", *Boston Review 31 (4). [MIT mimeograph, oct-2003]*

BUCHANAN-SMITH, M. (2003) "How the Sphere Project Came into Being: A Case Study of Policy making in the Humanitarian Aid Sector and the Relative Influence of Research", *ODI Working Paper* 215.

CAD (1991) *The DAC Principles for the Evaluation of Development Assistance*. OECD. París.

CAD (1998) *Review of the DAC Principles for evaluation of development assistance.* OECD. Paris.

CAD (2002) *Glosario de los principales términos sobre evaluación y gestión basada en resultados.* Evaluation and Aid Effectiveness Nº6. OECD-DAC. Paris.

CASLEY, D. y KUMAR, K. (1990) *Recopilación, análisis y uso de los datos de seguimiento y evaluación.* Mundi-Prensa y Banco Mundial. Madrid.

CHAMBERS, R. (1997) *Whose Reality Counts? Putting the first, last.* Intermediate Technology Publications. London.

DJEBBARI, H. & SMITH, J. (2005) "Heterogeneous Program Impacts of PROGRESA", Laval University and University of Michigan.

DUFLO, E. (2005) "Field Experiments in Development Economics", paper prepared for the World Congress of the Econometric Society, December.

DUFLO, E. & KREMER, M. (2005) "Use of Randomization in the Evaluation of Development Effectiveness", in PITMAN, G.; O. FEINSTEIN & G. INGRAM (eds) *Evaluating Development Effectiveness.* World Bank Series on Evaluation and Development. Vol.7. Transaction Publishers. New Brunswick.

EASTERLY, W. (2001) *The Elusive Quest of Growth.* MIT Press. Cambridge.

EASTERLY, W. (2006) *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. Penguin Press, New York.

FETTERMAN, D. (2001) *Foundations of Empowerment Evaluation.* Sage. Thousand Oaks.

GLEWWE, P; N. ILIAS; & M. KREMER (2003) "Teacher Incentives" *NBER Working Paper: No. 9671.*

GLEWWE, P; M. KREMER; S. MOULIN & E. ZITZEWITZ (2004) "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics 74, 251-268.*

GOMEZ GALÁN, M. y CÁMARA, L. (2003) *Orientaciones para la aplicación del enfoque del marco lógico. Errores frecuentes y sugerencias para evitarlos.* CIDEAL. Madrid.

GUSTAFSON, P. (2003) "How Random Must Random Assignment Be in Random Assignments Experiments?", *Social Research and Demonstration Corporation Technical Paper 03-01.*

HODDINOTT, J. & SKOUFIAS, E. (2004) "The Impact of PROGRESA on Food Consumption", *Economic Development and Cultural Change 53 (1), 37-61.*

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

KREMER, M. (2003) "Randomized Evaluations of educational Programs in Developing Countries: Some Lessons", *American Economic Review 93 (2), 102-106.*

KREMER, M. & MIGUEL, E. (2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", *Econometrica 72 (1), 159-217.*

KRUEGER, A. & WHITMORE, D.M. (2001) "The Effect of Attending Small Class in Early Grades on College Test-Taking and Middle School Test Results: Evidence from Project STAR," *The Economic Journal 111, 1-28.*

LALONDE, R. (1986) "Evaluating the econometric Evaluations of Training Programs with Experimental Data", *American Economic Review 76 (4), 604-620.*

LARRÚ, J.M. (2000) *La evaluación en los proyectos de cooperación al desarrollo de las ONGD españolas.* Tesis Doctoral. Universidad San Pablo-CEU. Madrid.

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

MAE (1998) *Metodología de Evaluación de la Cooperación Española.* MAE-SECIPI. Madrid.

MAE (2001) Metodología de Evaluación de la Cooperación Española II. MAE-SECIPI. Madrid.

MAYOUX, L. & CHAMBERS, R. (2005) "Reversing the Paradigm: Quantification, Participatory Methods and Pro-Poor Impact Assessment", *Journal of International Development 17 (2), 271-298.*

MORENO-DODSON, B. (ed.) (2005) *Reducing Poverty on a Global Scale. Learning and Innovating for Development. Findings from the Shanghai Global Learning Initiative.* The World Bank. Washington.

RAUDENBUSH, S. et al. (2006) "Optimal Design for Longitudinal and Multilevel Research : Documentation for the « Optimal Design » Software".

RAVALLION, M. (2005) "Evaluating Anti-Poverty Programs", *World Bank Policy Research Working Paper 3625.*

ROCHE, C. (2004) *Evaluación de impacto para agencias de desarrollo.* Cuadernos de Cooperación de Intermón-Oxfam. Barcelona.

SAVEDOFF, W.; R. LEVINE & N. BIRDSALL (2006) "When Will We Ever Learn? Improving Lives Through Impact Evaluation", Report of the Evaluation Gap Working Group. Center for Global Development, may.

SCHULTZ, T.P. (2004) "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program", *Journal of Development Economics 74 (1), 199-250.*

SKOUFIAS, E. (2005) *PROGRESA and Its Impact on the Welfare on Rural Households in Mexico.* Research Report 139, International Food Research Institute. Washington.

SCRIVEN, M. (1973) "Goal-Free Evaluation", HOUSE, E.R. (ed.) *School Evaluation. The politics and process.* McCutchan. Berkeley.

STOREY, A. (1997) "Non neutral humanitarism: NGOs and the Rwanda crisis", *Development in Pactice 7 (4), 384-394.*

STUFFLEBEAM, D.L. y SHINKFIELD, A.J. (1993) *Evaluación sistemática. Guía teórica y práctica.* Temas de educación Paidós./M.E.C. Barcelona.

UVIN, P. (1998) *Aiding Violence. The Development Enterprise in Rwanda.* Kumarian Press. West Hartford.

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.