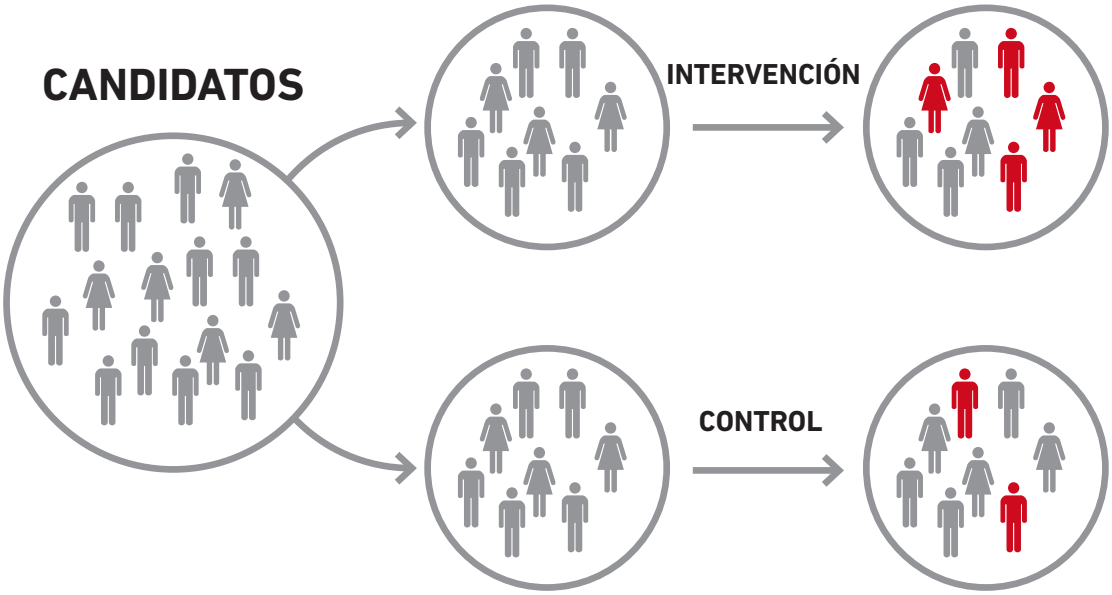
Guía práctica 10 - Evaluar el impacto de las políticas activas de ocupación

Colección Ivàlua de guías prácticas sobre evaluación de políticas públicas



ivàlua | Institut Català d'Avaluació de Polítiques Públiques

SOC Servei d'Ocupació de Catalunya

Institucions membres d'Ivàlua

Generalitat de Catalunya Departament d'Economia i Coneixement

Diputació Barcelona

upf. Universitat Pompeu Fabra Barcelona

Fundació Jaume Bofill

CTESC Consell de Treball Econòmic i Social de Catalunya

Generalitat de Catalunya Consell Interuniversitari de Catalunya

Autores:
David Casado,
Federico Todeschini

Maquetación y diseño portada: jaumebadosa.es

Primera edición: Junio 2013

Con la colaboración de:

Obra Social "la Caixa"

Depósito legal: B-22903-2013

ÍNDICE

1. INTRODUCCIÓN A LA GUIA

En un contexto como el actual, con unos presupuestos públicos lastrados por la crisis económica, los recursos destinados a financiar las políticas activas de empleo (PAE) han experimentado una disminución muy notable tanto en Cataluña como en el resto del Estado español. No obstante, los responsables políticos —independientemente de su color— insisten en la necesidad de «mejorar la efectividad» de los programas o de «ganar en eficiencia» en la prestación de los servicios. Esta insistencia nos parece positiva, aunque no creemos que se mantenga cuando vengan tiempos mejores, y esto sería un error: preguntarse si una determinada política activa aumenta la inserción laboral de los desempleados (efectividad) o si, en comparación con otras intervenciones, el coste de esta política por cada desempleado insertado es más o menos favorable (eficiencia) son cuestiones que la Administración debería plantearse en todo momento, sea cual sea la situación económica y el estado de las finanzas públicas. Si no se preocupa por ello, corremos el riesgo de seguir financiando programas que no generen los efectos buscados o dejar de financiar intervenciones que los consigan con creces.

A pesar de que, en términos relativos, por el momento se han evaluado más las PAE que otras políticas públicas, como las de educación o las de servicios sociales, aún nos encontramos muy lejos de los niveles que se observan en otros países de nuestro entorno. Por ejemplo, en una reciente revisión de las evaluaciones de impacto de políticas activas de empleo realizadas en Europa (Kluve, 2010) en la que se incluían únicamente aquellas que superaban un determinado umbral de rigor metodológico, solamente hay tres evaluaciones españolas de un total de 137. La situación en Cataluña es algo mejor, gracias a las evaluaciones impulsadas por el SOC durante el periodo 2007-2012, realizadas tanto por Ivàlua como por otras instituciones. No obstante, hay pocos motivos para la autocomplacencia, ya que la percepción por parte del principal financiador de las PAE —la Unión Europea— es que a España le queda un largo camino por recorrer a este respecto. Tanto es así que incluso el *Memorandum of Understanding* (MoU) menciona esta situación, solicitando a España en unos de sus puntos que impulse evaluaciones rigurosas sobre la efectividad de las PAE.

En definitiva, no solo es importante medir el impacto, sino que la presión para hacerlo, y para hacerlo bien, irá aumentando con rapidez. En este contexto, el principal propósito de esta guía es explicar a una audiencia no especializada, compuesta por políticos y gestores, los principales aspectos de la evaluación de impacto, haciendo un énfasis especial en su aplicación en el ámbito de las PAE[1]. En concreto, después de describir qué quiere decir evaluar el impacto o la efectividad de un programa (apartado 2), se describen los principales diseños que pueden utilizarse a tal efecto (apartado 3). Finalmente, se discuten los principales elementos que, a nuestro parecer, deberían tenerse en cuenta para impulsar un esquema estable de evaluaciones de impacto de las PAE (apartado 4). En este sentido, esta guía es un complemento del documento *Modelo de evaluaciones del SOC*, elaborado por Ivàlua recientemente[2].

Notas:

[1] *Los lectores interesados en profundizar en los aspectos más técnicos de la evaluación de impacto encontrarán al final de esta guía una selección de lecturas comentada.*

[2] *Esta guía se centra en la evaluación de impacto o de efectividad. Hay que tener presente, sin embargo, que existen otros tipos de evaluaciones, tales como la evaluación de necesidades, de la implementación, la evaluación ex ante o la evaluación económica, que exploran otras cuestiones igualmente importantes a la hora de valorar una política pública. Las guías 2, 4, 6 y 7 de la Colección Ivàlua de guías prácticas sobre evaluación de políticas públicas ofrecen, respectivamente, sendas introducciones al respecto. Estas guías están disponibles en http://is.gd/qWVvkR.*

2. QUÉ QUIERE DECIR EVALUAR EL IMPACTO Y CÓMO NO DEBE CALCULARSE

La mayoría de políticas públicas —y las PAE no son una excepción— pretenden en última instancia modificar algún *outcome* que se considera relevante. Por ejemplo: los programas del SOC cuyos destinatarios son los jóvenes desempleados buscan, mediante actuaciones diversas mejorar su grado de inserción laboral. En este sentido, si existiera, por ejemplo, un hipotético programa denominado Segunda Oportunidad que consistiera en proporcionar cuatro meses de formación profesional a jóvenes parados sin la ESO, ¿cómo podría valorarse si el programa mejora o no la inserción laboral?

Una aproximación muy sencilla para intentar responder a esta pregunta consiste en valorar la situación laboral de los beneficiarios del programa un tiempo después de haberlo terminado, al cabo de, por ejemplo, seis meses. Supongamos que en el caso de Segunda Oportunidad, por ejemplo, un 43 % de los beneficiarios hubiera encontrado trabajo para entonces. ¿Podemos afirmar que haber logrado este nivel de inserción es atribuible exclusivamente a la participación de los jóvenes en el programa? No, en absoluto. El hecho de que los jóvenes beneficiarios hayan encontrado trabajo o no puede depender de muchas cosas que no tienen nada que ver con el programa, tales como la situación del mercado laboral, la existencia de una política estatal que incentive la contratación juvenil o la intensidad a la hora de buscar trabajo, entre otros factores. De hecho, una vez que se reconoce esta multiplicidad de causas, la cuestión relevante es si la tasa de inserción laboral de los jóvenes de Segunda Oportunidad habría sido del 43 % si no hubieran participado en el programa: solamente en el caso de que fuera inferior, podríamos afirmar realmente que el programa ha tenido un «impacto» positivo sobre la inserción laboral de los beneficiarios de Segunda Oportunidad; de hecho, si la tasa de inserción en ausencia del programa fuera superior al 43 %, el impacto sería negativo.

En términos más generales, el impacto de una intervención o de un programa es la diferencia entre lo que realmente ocurre a los participantes y el denominado *contrafactual*: es decir, lo que les habría sucedido de no haber participado. Pero dado que es imposible que los mismos sujetos participen y no participen en un determinado programa, el *outcome* contrafactual no es observable. El principal reto de la evaluación de impacto, por tanto, consiste en proponer una estrategia que permita obtener una medida creíble de este *outcome* contrafactual; a continuación, si del *outcome* observado se resta esta estimación del *outcome* contrafactual, se obtiene una medida del impacto real del programa.

En este sentido, tal y como nos ha permitido ilustrar el ejemplo de Segunda Oportunidad, parece claro que utilizar las tasas de inserción laboral posprograma como medida del impacto de una PAE resulta completamente erróneo, salvo que estemos dispuestos a defender que el único factor que determina la inserción laboral de los desempleados son estas políticas (o, dicho con otras palabras, que ninguno de los beneficiarios del programa habría encontrado

trabajo si no hubiera participado en él). No obstante, la lista de procedimientos incorrectos para medir el impacto de una PAE no termina en el uso de la inserción laboral posprograma. Existen dos procedimientos más que, si bien reconocen la importancia de definir un contrafactual, conducen a estimaciones de impacto erróneas.

Una primera aproximación consiste en preguntar directamente a los beneficiarios, una vez finalizada su participación en el programa, si su empleabilidad ha mejorado o no gracias a ello. A este respecto, aunque en nuestro contexto no se realizan encuestas de este tipo, en otros países, como los Estados Unidos, constituyen un instrumento rutinario de seguimiento de la mayoría de las PAE (Smith, 2004). En el fondo, se espera de los individuos que estimen el impacto de su participación en el programa a partir de hacer más o menos explícita la idea de *outcome* contrafactual. Aunque a priori la idea de preguntar directamente a los beneficiarios puede resultar atractiva, la evidencia disponible sobre la robustez de los resultados de estas autoevaluaciones de impacto resulta decepcionante: cuando se comparan los resultados obtenidos mediante encuestas con los impactos estimados por los propios beneficiarios mediante evaluaciones experimentales —que, como veremos más adelante, constituyen la metodología más sólida para estimar impactos— la conclusión es que a los individuos no se les (nos) da bien evaluar impactos (Heckman y Smith, 1998). Este resultado no debería sorprendernos mucho: por una parte, los individuos tienen todo tipo de problemas cognitivos a la hora de tomar buenas decisiones sobre cuestiones mucho más sencillas que la de estimar contrafactuales, como ponen de manifiesto las numerosas investigaciones realizadas por los psicólogos y economistas del comportamiento; por otra parte, cuando los beneficiarios de un programa sospechan que sus respuestas pueden condicionar su continuidad en él, tienen incentivos a comportarse estratégicamente y sobrestimar sus valoraciones sobre qué impactos consigue el programa.

La segunda alternativa que se propone a menudo para medir el impacto de un programa pasa por comparar la inserción laboral de los participantes con un conjunto de no participantes similares en algunas características relevantes, como la edad o su nivel formativo. En el caso de Segunda Oportunidad, por ejemplo, podríamos comparar la inserción laboral a seis meses de los beneficiarios del programa (43 %) con la inserción correspondiente del resto de desempleados de Cataluña, también jóvenes y sin la ESO (30 %). Pero, ¿constituye este 30 % de inserción laboral un contrafactual creíble del *outcome* laboral de los participantes o, en otras palabras, el impacto del programa es de 13 puntos porcentuales?

La respuesta es, de nuevo, no, y el motivo tiene que ver con el **mecanismo de selección de beneficiarios** que utilizan las PAE, entre las que Segunda Oportunidad no tendría por qué ser una excepción. En particular, aunque nuestro hipotético programa habría estipulado los criterios básicos que deberían reunir los beneficiarios, como la edad, el tiempo en desempleo y la baja formación, hay dos «fuerzas» que gobiernan el proceso de participación en el programa y que van mucho más allá de los atributos que habrían aparecido en la convocatoria de Segunda Oportunidad.

Por una parte, entre todos los jóvenes desempleados sin la ESO que podrían participar en este programa que, como la mayoría de las PAE también sería de carácter voluntario, únicamente acabarían postulándose algunos de ellos y, en caso de ser escogidos, solamente algunos de ellos acabarían participando. La cuestión clave es que tanto la decisión de postularse inicialmente como la de querer participar en caso de ser aceptados están muy influidas por factores como la motivación, la renta familiar o la implicación de los padres, las cuales pueden ser muy diferentes de las del conjunto de jóvenes desempleados no participantes e incidir, seguramente, sobre las probabilidades de posterior inserción laboral de los individuos.

Por otra parte, además del comportamiento de los potenciales beneficiarios, la selección de participantes refleja también las decisiones que se toman desde el lado de la oferta, típicamente por parte de los gestores de las Oficinas de Empleo o de los técnicos de las entidades proveedoras de servicios. De nuevo, además de los criterios formales de cada convocatoria, estos profesionales escogen de entre el conjunto de desempleados «elegibles» no solamente a los que se informan sobre un determinado programa, sino también a los candidatos más idóneos de entre los que se acaban postulando, generalmente mediante la realización de entrevistas, dado que suele haber un exceso de demanda considerable. La cuestión importante es que en este proceso de selección hay, una vez más, múltiples atributos de los candidatos que serán tenidos en cuenta, como la motivación, la buena presencia, etc., los cuales también tienen una clara influencia sobre las posibilidades de encontrar un trabajo más adelante.

Así pues, fruto de la concurrencia de los dos procesos anteriores, los jóvenes desempleados con baja formación que acaben participando en Segunda Oportunidad pueden ser muy distintos del resto de desempleados que, a pesar de ser jóvenes y no tener la ESO, no acabará participando. Además, dado que estas diferencias influyen sobre las posibilidades de encontrar un trabajo, la comparación entre la tasa de inserción laboral de los participantes con la de los no participantes no puede considerarse una medida válida del impacto del programa, ya que una parte indeterminada de las diferencias de ambas tasas de inserción (43 % frente a 30 %) tendrá su origen en la distinta composición de ambos grupos de jóvenes, un fenómeno que se conoce con el nombre de *sesgo de selección*. Tal y como se explica en el siguiente apartado, los métodos que se utilizan para evaluar el impacto de una política intentan, mediante estrategias diversas, definir grupos de comparación que no estén expuestos a este sesgo de selección que acabamos de mencionar.

No obstante, antes de presentar este conjunto de diseños evaluativos, conviene dejar bien claro que la pretensión de medir correctamente el impacto del programa no es una mera fijación académica, propia de investigadores cegados por el rigor científico. El hecho de medir incorrectamente el impacto de un programa puede resultar muy contraproducente para los beneficiarios de las PAE, los cuales deberían ser, en última instancia, la preocupación fundamental de académicos, políticos y gestores. Supongamos, por ejemplo, que las decisiones

sobre asignación de recursos entre las distintas PAE, o entre proveedores de una determinada PAE, se llevaran a cabo de acuerdo con los niveles de inserción laboral posprograma. En ambos casos, a menos que las poblaciones atendidas fueran completamente equivalentes, una parte de las diferencias entre los niveles de los *outcomes* posprograma serán el resultado de diferencias en las características de los individuos (sesgo de selección) o se deberán a la existencia de factores contextuales diferentes. De hecho, sería perfectamente posible encontrar situaciones en las que niveles elevados de inserción laboral posprograma coexistieran con impactos nulos o negativos y casos en los que los niveles de inserción laboral fueran bajos, pero los impactos positivos. Así pues, si queremos acabar destinando más recursos a las políticas o a los proveedores que realmente «hacen más» por incrementar la empleabilidad de los desempleados, resulta fundamental que midamos correctamente los impactos. En la siguiente sección presentamos, tal y como habíamos avanzado, las principales técnicas que utiliza la evaluación de impacto para medir correctamente la efectividad de los programas.

3. DISEÑOS ALTERNATIVOS PARA EVALUAR IMPACTOS

El gran reto de la evaluación de impacto, como ya hemos señalado, radica en la estimación de un contrafactual creíble: una medida válida de lo que les habría ocurrido a los beneficiarios de la política si no hubieran participado. En este sentido, las PAE constituyen un campo de intervención pública propicio para la evaluación de impacto, ya que, dado que la mayoría de programas son voluntarios y su cobertura es limitada, suele ser posible disponer de un grupo potencial de individuos no participantes susceptible de actuar como grupo de comparación. Por otra parte, puesto que las experiencias piloto son habituales en este campo, de modo que algunos de los nuevos programas se prueban primero en un conjunto limitado de territorios, con frecuencia surgen oportunidades de encontrar territorios o personas que no han participado en la política que queremos evaluar. De hecho, como se explica en los apartados que siguen, lo que diferencia a las distintas técnicas de evaluación de impacto es, básicamente, la manera en que cada una de ellas define al grupo de comparación que se utiliza para estimar el contrafactual[3]. La exposición de las diversas técnicas va acompañada de sendos ejemplos ilustrativos —siempre procedentes de evaluaciones reales de PAE— que están recogidos y comentados en el anexo de esta guía.

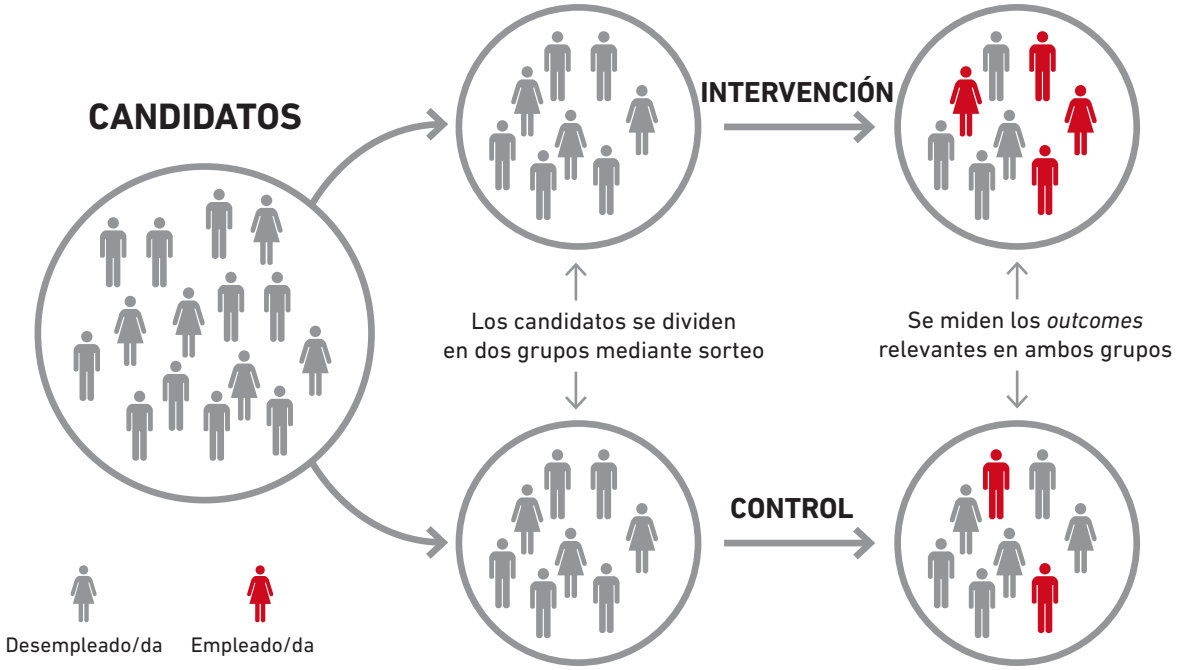
3.1 EVALUACIONES EXPERIMENTALES [4]

¿Què es una evaluación experimental?

Evaluar el impacto de una política pública mediante un diseño experimental es, desde una perspectiva metodológica, muy similar a aplicar la lógica que siguen los ensayos clínicos para testar la efectividad de un fármaco. La diferencia radica en que lo que evaluamos no es un fármaco, sino un programa, y que el *outcome* sobre el que buscamos producir un efecto no es el estado de salud, sino alguna problemática que afecta a los destinatarios.

Utilizaremos un ejemplo para ilustrar el funcionamiento de una evaluación experimental. Siguiendo con Segunda Oportunidad, la evaluación experimental de este hipotético programa podría realizarse de la manera siguiente: 1) dar instrucciones a las oficinas de empleo para que identifiquen a potenciales beneficiarios del programa, con el objetivo de alcanzar una cifra de mil candidatos; 2) mediante un procedimiento aleatorio, y previo consentimiento de los candidatos, aleatorizaríamos la participación en Segunda Oportunidad: 500 jóvenes lo recibirían y 500 no; y 3) transcurrido un cierto tiempo después de la finalización del programa, compararíamos los *outcomes* relevantes —como el grado de inserción laboral— entre el grupo de tratamiento y el de control. La figura 1 muestra gráficamente la esencia de una evaluación experimental del programa Segunda Oportunidad.

Figura 1.



Fuente: *Adaptado de Haynes et al. (2012).*

Ahora bien, ¿por qué la aleatorización, ya sea de pacientes en un ensayo clínico o de jóvenes desempleados en nuestro ejemplo, permite evaluar el impacto de un tratamiento o programa de una forma más válida que otras técnicas?

Gracias a la aleatorización, un experimento consigue que el grupo de tratamiento y el de control estén «equilibrados» en todos aquellos atributos personales que pueden influir sobre el *outcome* de interés, como puedan ser, en el caso de Segunda Oportunidad, la motivación, la experiencia laboral previa o el hecho de ser inmigrante o no serlo. De esta manera, cuando ya ha finalizado el programa y comparamos los *outcomes* entre los dos grupos para inferir el impacto, podemos descartar que el resultado obtenido sea la consecuencia de que los grupos son distintos; en otras palabras, la aleatorización nos ha permitido eliminar el sesgo de selección que mencionábamos en el apartado anterior. Por otra parte, dado que los dos grupos están expuestos a los mismos «factores de contexto» mientras dura el programa, como podría ser, por ejemplo, una mejora del mercado laboral en el caso de Segunda Oportunidad, también podemos descartar que estos factores sean los responsables de las diferencias postratamiento en los *outcomes*. En resumen, si detectamos estas diferencias en los *outcomes* entre ambos grupos, los podremos atribuir al único rasgo que los diferencia: haber participado o no en el programa. El experimento social nos proporciona, por tanto, una estimación válida del impacto del programa.

¿Qué se entiende exactamente por aleatorización?

La aleatorización de la participación constituye la piedra angular de un experimento social (ES) y su existencia es un requisito indispensable para que una evaluación pueda ser considerada experimental. La aleatorización que caracteriza a un ES no debe confundirse con el muestreo aleatorio que debe exigirse a una encuesta —ya sea de salud o de población activa— para que los resultados obtenidos sean representativos de la población. Por una parte, mientras que lo que debe ser aleatorio en una encuesta es la selección de los sujetos entrevistados, en un experimento social se exige que, entre los candidatos a participar en el programa, la elección de los que acaban participando y los que no se lleve a cabo mediante un procedimiento aleatorio. Por otra parte, la aleatorización en un ES no busca la representatividad de los resultados, sino permitir estimar sin sesgos el impacto del programa evaluado, como mencionamos anteriormente.

Esto no quiere decir, sin embargo, que un experimento social no pueda aleatorizar también el proceso de captación de candidatos. Por ejemplo, si los mil colegios de Cataluña estuvieran dispuestos a participar en un programa de incentivos a profesores, pero solamente hubiera presupuesto para aplicar el esquema en 100 centros, podríamos escoger 200 colegios al azar de entre los mil (muestra de candidatos) y, a continuación, asignar aleatoriamente la participación en el programa a la mitad de ellos. La primera aleatorización conferiría «representatividad» a nuestros resultados, en el sentido de que podrían considerarse extrapolables a los 800 colegios «no experimentales», pero es la segunda aleatorización la que nos permite medir el impacto del programa y la única necesaria para definir una evaluación como experimental.

La evaluación experimental de las PAE

En el mundo anglosajón, especialmente en los Estados Unidos, la evaluación experimental de políticas públicas es una práctica frecuente desde hace más de 30 años y, de hecho, los programas laborales son uno de los ámbitos más activos al respecto, con decenas de intervenciones evaluadas experimentalmente. En Europa, en cambio, las evaluaciones experimentales de PAE eran una rareza hasta hace poco más de una década. Sin embargo, durante los últimos años se han empezado a realizar en Europa numerosas evaluaciones experimentales en el ámbito de las PAE, como la Employment Retention and Advancement (ERA) Demonstration realizada en el Reino Unido (Hendra et al. 2011) o la evaluación de diversos esquemas de activación de desempleados realizados en Dinamarca (Graversen y Van Ours, 2008), por poner un par de ejemplos. Una mención especial merece el Fonds d'Expérimentation pour la Jeunesse, que comenzó en Francia en el año 2008 y que, a través de una dotación de más de 200 millones de euros para el periodo 2009-2014, intenta favorecer la evaluación experimental de nuevas formas de intervención destinadas a combatir el fracaso escolar y la exclusión laboral de los jóvenes franceses. Los programas evaluados hasta ahora —o en curso de evaluación— son muy variados e incluyen desde intervenciones orientadas a incrementar la implicación de los padres en el proceso educativo hasta la provisión de tutorías

para prevenir el abandono escolar, pasando por la prestación de servicios de orientación laboral a jóvenes desempleados. Hay que decir que, hasta donde llega nuestro conocimiento, todavía está por nacer la primera evaluación experimental de una PAE en el Estado español.

Críticas a los experimentos

Un primer aspecto que suele criticarse de los experimentos es que son caros. No cabe duda de que esta apreciación se debe en parte al precedente sentado por los primeros ES realizados en los Estados Unidos, que precisaron de equipos muy numerosos, encuestas muy costosas, etc. Hoy en día, la informatización de los registros administrativos ha permitido esquivar en muchos casos la necesidad de realizar costosas encuestas sin perder por ello riqueza analítica, ya que la exhaustividad y fiabilidad de estos registros es muy notable. En este sentido, hay que mencionar que las evaluaciones experimentales de PAE europeas —como es el caso de las dos mencionadas anteriormente— utilizaban únicamente registros administrativos.

En cualquier caso, más allá de las consideraciones económicas, una de las críticas más frecuentes hacia las evaluaciones experimentales tiene un trasfondo ético: resulta inadecuado privar a determinados individuos (los del grupo de control) de los beneficios que supone una nueva política utilizando un mecanismo tan arbitrario como la aleatorización. La réplica de los que ven en los experimentos sociales una herramienta adecuada de evaluación se apoya en tres consideraciones.

En primer lugar, la presunción de que se está privando a algunos individuos de algo beneficioso no debería tener sentido si el experimento está justificado, ya que es precisamente la ausencia de datos sobre la efectividad del programa lo que justifica su evaluación. En segundo lugar, son pocas las ocasiones en que pertenecer al grupo de control implica no recibir ningún tipo de intervención, sino que lo que se compara es la nueva política respecto a «seguir como hasta ahora». Y en tercer lugar, hay situaciones muy frecuentes en las que la aleatorización puede considerarse un criterio de asignación equitativo como, por ejemplo, cuando la falta de recursos no permite atender de una sola vez a toda la población potencialmente beneficiaria de la política; de hecho, cuando se producen situaciones de este estilo, un diseño experimental más aceptable que utilizar una simple lotería entre individuos es optar por un despliegue aleatorizado, es decir, aleatorizar el momento del tiempo en el que distintos grupos de individuos o territorios comenzarán a recibir el nuevo programa.

Una última objeción que suele plantearse a los experimentos es que a menudo carecen de validez externa o, en palabras menos técnicas, que los resultados que se obtienen en lo que respecta el impacto de una política, aunque sí son válidos respecto a los sujetos, momento y lugar en que se realizó el experimento, pueden no ser extrapolables a otros contextos distintos. Los que se dedican a la experimentación social han intentado mitigar la falta de validez externa por dos vías. En primer lugar, son habituales las denominadas *evaluaciones* multi-site, en las

que el programa se evalúa aplicándolo en diferentes lugares (por ejemplo, varios municipios), con el objetivo de analizar hasta qué punto los resultados de impacto varían en función de los contextos. Por otra parte, cuando el número de réplicas experimentales de un determinado tipo de programa es suficientemente importante, puede llevarse a cabo lo que se denomina *metanálisis de los resultados obtenidos*, es decir, un ejercicio cuantitativo de síntesis que pretende establecer si el programa resulta efectivo con carácter general, con independencia de en qué poblaciones, lugares y momentos se aplique.

Al margen de las críticas anteriores, es evidente que las evaluaciones experimentales plantean retos importantes tanto desde una perspectiva metodológica como desde un punto de vista de viabilidad política. Sin embargo, tal y como hemos expuesto en otros lugares (Casado, 2012) donde se discuten con detalle estas otras cuestiones, lo que sorprende es que este diseño evaluativo, que se utiliza en otros países ampliamente y que es considerado el *gold standard* de la evaluación de impacto, no se haya utilizado todavía en el Estado español para evaluar una política pública, incluidas las PAE.

El resto de métodos de evaluación de impacto, que trataremos a continuación, comparten con el diseño experimental la definición de un grupo de comparación, integrado por no participantes, que también se utiliza para estimar el *outcome* contrafactual de los participantes. La existencia de un grupo de comparación, como ya hemos discutido en el caso de los experimentos, permite eliminar de la estimación de impacto la influencia de los denominados *factores contextuales* (desempleo general, otras PAE, etc.), ya que tanto los beneficiarios del programa como el grupo de comparación están igualmente expuestos a ellos.

En cambio, a diferencia del diseño experimental, que elimina el sesgo de selección mediante la aleatorización de la participación, el resto de métodos únicamente consiguen este objetivo si se cumplen ciertos supuestos relativos al mecanismo que gobierna la participación en el programa. En esencia, mientras que las evaluaciones experimentales eliminan el riesgo que supone el sesgo de selección alterando «por diseño» el proceso de captación de participantes, el resto de métodos aspiran a obtener medidas del impacto no contaminadas por este sesgo de la mano de la modelización estadística.

3.2 REGRESIÓN DISCONTÍNUA

La denominada regresión discontinua (RD) es probablemente la técnica de evaluación de impacto no experimental que ha tenido un crecimiento más intenso en los últimos años, especialmente en el ámbito educativo (Schlotter *et al*., 2010). Esta técnica se aplica cuando la participación en el programa que se pretende evaluar depende del hecho de que una variable tome un conjunto de valores determinados. Por ejemplo, volviendo al hipotético programa Segunda Oportunidad del que hemos hablado antes, podría darse el caso de que los requisitos de la convocatoria estipulasen que únicamente los individuos con una renta familiar menor a un

cierto punto de corte (por ejemplo, veinte mil euros) pueden participar en él. O bien podríamos pensar en la existencia de un baremo (por ejemplo, los individuos obtienen puntos en función del tamaño del hogar, el nivel educativo del cabeza de familia, la edad, etc.) que determina que solamente pueden participar en el programa los que obtienen una puntuación superior (inferior) a un determinado umbral preestablecido en la convocatoria. En cualquier caso, para poder implementar la técnica de la RD no importa si el punto de corte (de la renta familiar o del baremo) es más grande o más pequeño, sino que lo único importante es que haya un punto de corte.

La lógica del método de la regresión discontinua es la siguiente. El hecho de que la participación esté determinada por un punto de corte comportará que haya un salto, justamente en el umbral, en el porcentaje de individuos que participan del programa. De hecho, si el umbral opera de manera estricta, lo que observaremos alrededor del punto de corte serán dos grupos de individuos: los que participan en el programa y los que no. En nuestro ejemplo, si suponemos que el umbral se ha establecido en veinte mil euros, observaríamos un «salto» en el porcentaje de participantes justo por encima y por debajo de esta cantidad. Pues bien, si el programa Segunda Oportunidad resultara efectivo, entonces también esperaríamos que el salto en el porcentaje de participación se reflejara en la tasa de inserción laboral y, por tanto, hubiera también un «salto» entre la tasa de empleo de los individuos con una renta familiar que estuviera justo por encima y por debajo del umbral. La idea subyacente de la RD es que la cantidad exacta en la que se fija el umbral, ya sean 20 000, 10 000 o 15 000 euros, resulta básicamente arbitraria. Así pues, desde la perspectiva de los desempleados, el hecho de participar o no en el programa acaba dependiendo de algo que se parece a un «sorteo», ya que al final acceden o no en función de que la regla fijada por la Administración les deje por encima o por debajo del umbral. Y, si esto es cierto, hay que pensar que los que acaban participando y no participando, justo alrededor del umbral, son esencialmente idénticos en todos los atributos relevantes. De esta manera, se habría eliminado el sesgo de selección.

Sin embargo, como ya hemos mencionado antes, la técnica de la RD solamente proporciona estimaciones válidas si se satisface un conjunto de supuestos que hay que tener muy presente. En primer lugar, para que el método funcione es necesario suponer que no existe ninguna otra variable que «salte» en el mismo punto de corte. Por ejemplo, no podríamos aplicar el método de regresión discontinua si el punto de corte de veinte mil euros coincidiera con el punto de corte de otro programa o política (cursos de formación, pensión no contributiva, etc.), ya que si así fuera, no podríamos saber qué parte del salto en la tasa de empleo se debe a Segunda Oportunidad y qué parte al otro programa que también muestra un salto. Otra condición importante es que los individuos no puedan escoger en qué lado del punto de corte se encuentran. Por ejemplo, la estimación quedaría invalidada si los individuos más motivados pudieran reducir la renta de 20 050 euros a 19 950 euros con el objetivo de poder participar en Segunda Oportunidad. En este caso, no podríamos saber qué parte del salto en la tasa de empleo es debida al programa, y qué parte es debida a la distinta motivación que tienen los individuos cuya renta familiar es de 20 050 euros comparada con la de los individuos cuya renta familiar es de 19 950 euros.

En cualquier caso, aunque el método de la regresión discontinua proporciona buenas estimaciones del impacto cuando se satisfacen los supuestos anteriores, hay que tener presente que esta técnica solo informa sobre la efectividad del programa para aquellos individuos que están cerca del umbral. Así, siguiendo con nuestro ejemplo, el método no proporcionaría estimaciones válidas sobre el impacto del programa en el caso de individuos cuya renta familiar fuera de, por ejemplo, 30 000 o 10 000 euros, dado que están lejos del punto de corte de 20 000 euros que garantiza la similitud entre participantes y no participantes sobre la que descansa la bondad del método.

Finalmente, tenemos que decir que la aplicación de la RD en el ámbito de las PAE ha sido menos importante, aunque existen algunas excepciones notables, como es el caso de la evaluación del New Deal for Young People (Di Giorgi, 2008) que aparece resumida en el anexo. La principal razón de este uso limitado debemos buscarla, en opinión de Card *et al*. (2011), en el hecho de que la participación en la mayoría de PAE no suele venir definida por una regla de elegibilidad estricta y, cuando es así, o bien se tiende a aplicar de manera laxa, o los candidatos la conocen e intentan alterar su puntuación para poder acceder al programa (manipulando su nivel formativo, de renta, etc.). En cualquier caso, dado que la RD es una metodología que permite obtener estimaciones de impacto muy robustas cuando los supuestos en los que se basa se satisfacen, merece la pena tenerla presente y evaluar si puede aplicarse a cada caso en concreto.

3.3 MATCHING

Esta técnica trata de imitar un experimento mediante la definición a posteriori, con la ayuda de la estadística, de un grupo de comparación de no participantes que se parezca lo más posible a los beneficiarios del programa que se pretende evaluar. La aplicación de este método puede considerarse en los casos en que, después de haber terminado el programa, disponemos de información para los participantes y los no participantes sobre: 1) los *outcomes* respecto a los cuales queremos medir el impacto del programa, tales como la inserción laboral en el caso de un curso de formación ocupacional; y 2) todos aquellos factores (personales, familiares, de entorno, etc.) que simultáneamente hayan podido influir, por una parte, sobre el hecho de acabar participando o no en el programa y, por otra, sobre los *outcomes* respecto a los que queremos evaluar el impacto (por ejemplo, inserción laboral).

Lo que propone el método del *matching* es utilizar toda la información anterior para construir un grupo de comparación legítimo, constituido únicamente por no participantes que realmente se parezcan a los beneficiarios, para poder así estimar el impacto del programa simplemente como la diferencia entre los *outcomes* de ambos grupos cuando haya finalizado el programa. Para lograrlo, el método busca para cada uno de los participantes una pareja o match que sea lo más parecida posible en el conjunto de variables (personales, familiares y de entorno) mencionadas previamente. A tal efecto, pueden utilizarse algoritmos estadísticos diversos,

cuyos detalles conviene dejar al margen para mantener el tono divulgativo que hemos querido utilizar en este documento[5].

El supuesto básico que la técnica del *matching* necesita para obtener estimaciones válidas del impacto es que, más allá de las variables que se hayan tenido en cuenta para realizar los emparejamientos, no hay ningún otro factor que tenga influencia simultáneamente sobre el proceso de participación en el programa y sobre los *outcomes* de interés. O, dicho con otras palabras, la hipótesis fundamental es que todo el sesgo de selección de la participación en el programa se produce en variables sobre las que tenemos información, de tal manera que al tenerlas en cuenta a la hora de construir el grupo de comparación, la simple diferencia de *outcomes* entre participantes y no participantes «emparejados» constituye una medida válida del impacto del programa. Lógicamente, cuanto más amplio sea el conjunto de variables sobre las que se dispone de información, más plausible resultará defender que hemos conseguido reducir el problema del sesgo de selección.

De hecho, fruto de la riqueza y exhaustividad de las bases de datos disponibles en el ámbito de las PAE, no es raro que el *matching* se haya utilizado profusamente en otros países para evaluar el impacto durante estos últimos años (Kluve, 2010). Sin embargo, no son pocos los analistas que consideran que la identificación basada únicamente en variables observables resulta demasiado débil de cara a eliminar el sesgo de selección, ya que los procesos de captación de participantes de muchas PAE se caracterizan, precisamente, por un cribado sustancial basado en elementos actitudinales o motivacionales, completamente ausentes de las bases de datos. A pesar de ello, como apunta Vera (2012), algunos investigadores señalan que utilizar información previa sobre el *outcome* de interés (por ejemplo, la participación laboral), para diversos periodos anteriores al inicio del programa, puede ayudar mucho a obtener resultados fiables. En cualquier caso, al igual que ocurre con la regresión discontinua, la plausibilidad de los supuestos del *matching* deberán ser valorados a la luz de cada aplicación concreta.

3.4 MODELO DE DOBLES DIFERENCIAS

El modelo de dobles diferencias (DD) puede utilizarse en aquellas situaciones en las que un grupo de agentes (personas, OTG, municipios...) es expuesto a un determinado programa («tratamiento») y el resto no. El caso más sencillo es el de dos grupos y dos periodos: en el primer periodo, ninguno de los dos grupos participan del programa; en el segundo periodo, en cambio, un grupo participa y el otro no. Un ejemplo de una situación de este estilo, evaluada de hecho mediante un diseño DD, es la que acompañó a la implementación del New Deal for Young People en el Reino Unido, ya que el programa únicamente estuvo operativo durante la fase piloto en un conjunto limitado de territorios (Blundell et al., 2004).

La estrategia de identificación del impacto que propone un modelo de DD consiste en comparar, por una parte, la diferencia en los *outcomes* entre los dos grupos en algún momento después

de la finalización del programa y, por otra parte, la diferencia existente antes de iniciarlo. Una vez hecho esto, si el inevitable sesgo de selección entre ambos grupos permanece constante a lo largo del tiempo, el cambio entre antes y después en la diferencia de *outcomes* entre el grupo de participantes y el de comparación (es decir, la diferencia de las diferencias) constituye una estimación válida del impacto del programa.

El supuesto clave en un diseño DD es que cualquier diferencia preprograma entre el grupo de tratamiento y el de comparación se habría mantenido constante aunque el programa no hubiera existido. Bajo este supuesto, los *outcomes* posprograma del grupo de comparación, debidamente ajustados por el diferencial preprograma, se convierten en un contrafactual creíble de los *outcomes* posprograma del grupo de tratamiento. Es importante darse cuenta de que, comparado con el método del *matching*, el supuesto de un modelo de DD es menos restrictivo en la medida en que ambos grupos pueden ser diferentes en factores inobservables (la motivación, por ejemplo, en el caso de un programa de formación), siempre y cuando estas diferencias permanezcan constantes a lo largo del tiempo; en cambio, en el caso del *matching*, para que la técnica proporcione una medida válida del impacto, es necesario que ambos grupos no difieran en ningún factor no observable. No son pocos los investigadores, como Card et al. (2011), que ven en este punto una clara ventaja del modelo de DD sobre el *matching*, hasta el extremo de que recomiendan con entusiasmo la utilización del primero y con mucha más cautela la del segundo.

No obstante, no debemos olvidar que también el modelo DD, como el resto de diseños experimentales, está sujeto a limitaciones si no se satisfacen los supuestos que lo sustentan. En este sentido, el punto débil de los modelos de DD es la posible existencia de factores inobservables que varíen a lo largo del tiempo. Así, siguiendo con el ejemplo del programa Segunda Oportunidad, si la motivación de tratamientos y controles varía a lo largo del tiempo, y no podemos observar esta variable, no podremos estar plenamente seguros de que este factor no sea la causa de la evolución diferencial del *outcome* en el grupo de tratamiento respecto al de control y, por tanto, de que la magnitud del impacto estimado para la política no sobrestime su efecto real. Por consiguiente, si queremos que resulten creíbles los resultados de una evaluación de impacto que utilice un diseño DD, tendremos que presentar argumentos que permitan descartar la existencia de características inobservables que varíen en el tiempo de forma distinta entre tratamientos y controles.

3.5 ELECCIÓN ENTRE MÉTODOS

Los apartados anteriores han puesto de manifiesto la existencia de diversos métodos susceptibles de ser utilizados a la hora de intentar establecer el impacto de una determinada política. En general, una visión muy compartida entre los evaluadores es que no existe el método ideal, es decir, un tipo de diseño particular que, independientemente de las circunstancias, debería aplicarse de forma universal en todas las evaluaciones de impacto. En la práctica, por tanto, los

evaluadores se ven obligados a escoger entre varias alternativas. Un elemento obvio que condiciona estas elecciones es la disponibilidad de tiempo y de recursos, aunque también hay otros: las características del programa, la importancia de los resultados y el uso que se quiera hacer de ellos, la disponibilidad de datos, etc. Los apartados que siguen tratan brevemente sobre estos aspectos, y argumentan a favor de la necesidad de aproximarse a la elección del método con una mentalidad abierta, ecléctica y desprovista de apriorismos excesivos.

Hay determinadas características de las políticas públicas que aumentan las posibilidades de medir su impacto con rigor. Una especialmente importante es la relativa a su novedad y, más concretamente, a su concepción como prueba piloto. En estos casos, si se reúnen una serie de condiciones, como que la demanda potencial sea superior a la oferta o que existan dudas sobre la efectividad del programa, los experimentos sociales que utilizan procedimientos de asignación aleatorios pueden constituir una forma de evaluación de impacto a considerar. En cualquier caso, a pesar de que la asignación no se produce de manera aleatoria, un programa piloto que se implante solamente en determinadas zonas geográficas abre las puertas a diseños no experimentales (*matching* o modelos DD) que utilicen las áreas no piloto para construir grupos de comparación.

Otra ventaja de las políticas nuevas —se materialicen o no mediante pruebas piloto— es que permiten la introducción de elementos de evaluabilidad mientras se desarrolla la fase de diseño del programa. Como hemos mencionado anteriormente, una evaluación de impacto es, por definición, una evaluación *ex post*, pero las mejores evaluaciones de impacto son aquellas que se planifican *ex ante*. La posibilidad más extrema es que el propio despliegue de la política se realice pensando en la evaluación, como es el caso de un experimento social, pero a veces basta con planificar una buena recogida de datos antes y después de la intervención, que afecte a sendas muestras de potenciales beneficiarios y no beneficiarios, para incrementar enormemente las posibilidades de obtener estimaciones de impacto creíbles mediante métodos no experimentales.

Sin embargo, a menudo sucede que el impacto que se desea evaluar no es el de una política nueva. En estos casos, dado que resulta imposible influir en «clave evaluadora» sobre el diseño del programa, el reto de la evaluación consiste en encontrar características de la política y fuentes de información que hagan posible la aplicación de las técnicas cuasiexperimentales descritas en esta guía.

Así pues, en lo que respecta a las características del programa, hay que buscar elementos que posibiliten la construcción de contrafactuales: por ejemplo, si por los motivos que sea un determinado programa tiene listas de espera, los individuos incluidos en ella pueden constituir un grupo de control natural respecto del que estimar el impacto del programa. Asimismo, en la medida en que exista variabilidad geográfica en el grado de implantación de una política, las unidades territoriales que dispongan del programa pueden compararse con las que no lo

tienen (los municipios catalanes pueden constituir, en el caso de algunas PAE, una fuente de variabilidad a explorar en este sentido).

Las distintas técnicas en que hemos centrado nuestra atención hasta el momento son metodologías de análisis cuantitativas. No es extraño que este tipo de enfoque prevalezca en la evaluación de impacto, ya que la cuestión fundamental a resolver —que no es otra que la construcción de un contrafactual— es de naturaleza básicamente cuantitativa. A pesar de ello, existe la percepción creciente entre los evaluadores de que, con el fin de mejorar la robustez de la evaluación de impacto, resulta recomendable complementar el análisis utilizando **técnicas cualitativas** (entrevistas en profundidad, grupos de discusión, etc.)[6]. El valor añadido que puede aportar su utilización es permitir al equipo evaluador mejorar su conocimiento sobre las condiciones en que realmente opera el programa, las perspectivas de sus beneficiarios, así como otros elementos fundamentales a la hora de entender realmente el porqué del impacto de una política o programa (o de su ausencia).

Notas:

[3] *Hacemos abstracción de la medida del outcome en sí, que puede ser la inserción laboral u otras cosas, para concentrarnos en los aspectos relacionados propiamente con la medida del impacto. En el documento Modelo de evaluaciones del SOC, se tratan diversas cuestiones relativas a la medida de los outcomes en el caso de las PAE.*

[4] *Véase Casado (2012) para obtener un análisis más detallado de las posibilidades y retos que supone la evaluación experimental de políticas públicas. Por otra parte, para conocer diversas iniciativas orientadas a desarrollar institucionalmente este tipo de evaluaciones, consulte la panorámica de Lázaro (2012) al respecto.*

[5] *Los lectores interesados en profundizar en los aspectos más técnicos del matching pueden consultar las referencias que se mencionan en la sección «Para obtener más información» que cierra la guía.*

[6] *Véase Sanz (2011) para una revisión de las principales técnicas de análisis cualitativo y su aplicación a la evaluación de políticas públicas.*

4. ¿CÓMO MEJORAR LA EVALUACIÓN DE LAS PAE AHORA Y AQUÍ?

4.1 DISEÑAR LA POLÍTICA AL MISMO TIEMPO QUE LA EVALUACIÓN

El momento en que se diseña la evaluación de una política suele tener consecuencias importantes sobre la robustez de los resultados de la evaluación. En este sentido, a pesar de que toda evaluación de impacto mide retrospectivamente los efectos de un programa, algunas de ellas tienen la característica de haber sido diseñadas antes de implementar el nuevo programa.

En general, las evaluaciones de impacto diseñadas antes del inicio del programa, que denominaremos *prospectivas*, permiten aportar pruebas más rigurosas sobre los efectos de un programa por dos motivos fundamentales: por un lado, abren la posibilidad de influir sobre cómo aplicar el programa y, de esta manera, apostar por fórmulas de implementación que generen grupos de tratamiento y comparación que permitan estimar el impacto de forma más creíble (por ejemplo, llevando a cabo una evaluación experimental con aleatorización); por otro lado, ofrecen la posibilidad de definir desde el inicio las piezas de información que harán falta para realizar la evaluación (medidas adecuadas de los *outcomes*, de las características de los individuos, de los costes de implementación) y poner los medios para generarlas. En cambio, si la evaluación es diseñada cuando un programa ya se encuentra en marcha (o, muy a menudo, finalizado), la elección del analista se encuentra muy condicionada no solo por cuáles son los datos disponibles y su calidad, sino también por las características de implementación del programa que, en el peor de los casos, pueden incluso hacer imposible aplicar alguna de las técnicas mencionadas previamente al no existir la posibilidad de definir un grupo de comparación.

Hasta ahora, todas las evaluaciones de impacto de los programas del SOC que se han realizado, incluidas las de Ivàlua, han sido diseñadas con los programas en cuestión ya finalizados. Así pues, uno de los retos del SOC es incorporar la evaluación desde el inicio del proceso de gestación de una nueva PAE. Esto pasa por describir claramente qué métodos de estimación contrafactual se utilizarán (experimento social, dobles diferencias, etc.), los *outcomes* concretos sobre los que se medirá el impacto (participación laboral, ingresos, etc.) y las fuentes de información que se utilizarán no tan solo para medir los *outcomes*, sino también el resto de variables mencionadas previamente (características de los individuos, costes de implementación, etc.).

De hecho, más allá de que las evaluaciones prospectivas permitan estimar el impacto de los programas de forma más rigurosa, el mero ejercicio de «pensar en cómo evaluar el impacto» de un programa nuevo ya tiene efectos positivos sobre el propio diseño sustantivo de la intervención. En concreto, obliga a los responsables del programa a precisar nítidamente cuáles son los *outcomes* de interés sobre los que la intervención pretende tener algún impacto,

así como los mecanismos por los que se espera que estos efectos se produzcan. Además, al afrontar la necesidad de sustantivar *outcomes*, mecanismos e impactos esperados, los encargados de diseñar el programa tienden a revisar más el contenido y los impactos de las políticas que se han puesto en marcha en otros países, lo cual resulta positivo en la medida en que se aprovecha el conocimiento acumulado en otros lugares.

El fomento de la realización de evaluaciones prospectivas comienza a impregnar las actividades de instituciones diversas. Es el caso, por poner un ejemplo, del Banco Iberoamericano de Desarrollo (BID), que ha incorporado entre los ítems de priorización de los proyectos susceptibles de recibir financiación la obligatoriedad de diseñar la evaluación de impacto de la intervención. Algunas iniciativas recientes de la Comisión Europea, tales como las convocatorias PROGRESS de impulso a la realización de evaluaciones de impacto de las PAE, denotan un claro interés en el entorno europeo por avanzar en esta misma dirección.

4.2 LAS EXPERIENCIAS PILOTO: NO MÁS OCASIONES PERDIDAS

Las denominadas *experiencias piloto* representan un ámbito en el que la realización de evaluaciones prospectivas se convierte, a priori, en algo especialmente fácil de impulsar. En el caso de las PAE —y también en el de otras intervenciones educativas o de los servicios sociales—, no resulta extraño en nuestro país proponer una implementación a pequeña escala de un nuevo programa o política con el propósito precisamente de comprobar si se trata o no de una buena idea. Sin embargo, salvo excepciones puntuales, la mayoría de estas experiencias piloto no han definido una estrategia evaluativa que permita estimar con un mínimo de rigor el impacto de la intervención en cuestión (esto es, proponiendo el uso de algunas de las técnicas descritas en el apartado 3 de esta guía). En general, parece que el principal interés de estas experiencias piloto es someter a prueba elementos relativos a la implementación del programa —incluida su viabilidad «política»— con el propósito de mejorar su diseño antes de proceder a su generalización, la cual, desafortunadamente, a menudo ya está decidida antes de iniciar la prueba piloto.

No obstante, la mayoría de pilotos presentan, por naturaleza, algunas características muy propicias para la realización de evaluaciones prospectivas. En primer lugar, dado que por definición un piloto solamente se aplica a una fracción de los potenciales beneficiarios —ya sean personas, entidades o municipios— la existencia de un grupo de comparación está siempre garantizada. De hecho, dado que todo el mundo acepta que si se lleva a cabo un piloto habrá potenciales beneficiarios que no participarán en el programa, se abre la puerta a utilizar la aleatorización como mecanismo de asignación, ya que el azar suele percibirse en estas situaciones como un criterio de selección justo. En definitiva, si el SOC, por ejemplo, emprendiera un piloto para probar la efectividad de un nuevo dispositivo de asesoramiento para los desempleados, y hubiera presupuesto para financiar su implementación solamente en 15 OTG, podría optarse por seleccionar aleatoriamente las OTG participantes y de este modo evaluar experimentalmente el impacto de la intervención.

Por otra parte, aunque no se aleatorice cuáles de los potenciales beneficiarios se convertirán en participantes, la lógica subyacente del piloto es que se está «probando» algo y, por tanto, esto facilita la introducción de consideraciones evaluativas en el diseño del programa. En particular, y continuando con el ejemplo anterior, si la previsión fuera introducir el nuevo dispositivo en 15 OTG no escogidas al azar, todavía sería posible diseñar de antemano una evaluación de impacto basada en un modelo de dobles diferencias que comparase antes y después la evolución de los *outcomes* de interés en las 15 OTG participantes y en las 56 que no participan. El diseño de la evaluación incluiría una previsión de toda la información que habría que recopilar con el fin de tener el máximo de posibilidades de obtener una estimación robusta del impacto del programa.

4.3 EL EXCESO DE DEMANDA Y LA ALEATORIZACIÓN COMO MECANISMO DE ASIGNACIÓN JUSTO

Es poco habitual que los responsables de una determinada PAE —ya sean los planes de empleo, la formación ocupacional o el programa Suma't— reconozcan abiertamente la existencia de un exceso de demanda. Se trata de un posicionamiento comprensible, dado que el exceso de demanda es interpretado automáticamente por la opinión pública como una señal de insuficiencia de recursos. No obstante, a pesar de que es cierto que el exceso de demanda solamente debería preocuparnos en la medida en que contenga personas que necesitan el programa, cuesta creer que con tasas de cobertura como las que se observan en algunos programas, por debajo del 10 % de los desempleados elegibles (Suma't o NCNO), no haya desempleados que, a pesar de necesitar el programa, no accedan a él a causa de la restricción de plazas.

La reacción por parte de las entidades proveedoras de PAE con exceso de demanda es implantar estrictos procesos de selección entre los candidatos elegibles: así pues, de entre todos aquellos que reúnen los requisitos formales de la convocatoria, solamente acaban participando los que además demuestran estar más motivados, parecen tener más aptitudes, etcétera. En este sentido, sin entrar a discutir si realmente estos criterios de selección son los más adecuados o no, lo que proponemos es hacerlos explícitos, mediante sistemas de baremación que, de hecho, ya utilizan informalmente muchas de las entidades, y realizar sorteos cuando haya varios candidatos empatados en puntos.

Una alternativa más radical, en la que las entidades proveedoras perderían buena parte de su discrecionalidad actual, sería hacer un sorteo entre todos los desempleados que, por una parte, manifestasen interés en participar en el programa y, por otra, reunieran los requisitos previstos de la convocatoria. La asignación aleatoria de la participación podría considerarse un criterio de priorización justo si, como sería deseable, los criterios de la convocatoria capturasen adecuadamente la necesidad (o intensidad de la problemática) de los desempleados a los que el programa pretende ayudar.

4.4 LOS SUPLENTES: HACIENDO DE LA NECESIDAD VIRTUD

La existencia de suplentes es habitual en la mayoría de programas del SOC que tienen un componente formativo. Este es el caso, lógicamente, de la formación ocupacional (FOAP), pero también de otros programas donde se contemplan acciones formativas de naturaleza diversa, como es el caso de Suma't o de Noves Cases per Nous Oficis, por poner tan solo un par de ejemplos. Los suplentes se caracterizan por reunir los requisitos de la convocatoria y por haber superado el proceso de selección adicional al que nos hemos referido antes, aunque finalmente no han iniciado la participación en el programa porque había otros candidatos más adecuados o que habían llegado un poco antes. Sin embargo, si finalmente se producen bajas entre los titulares, se llamará a algunos de los suplentes para que se conviertan en participantes.

El hecho de que haya suplentes se explica por la concurrencia simultánea de dos factores. En primer lugar, son el resultado de la existencia de un exceso de demanda que afecta a personas que lo necesitan, les interesa y podrían obtener un provecho, ya que los propios gestores del programa aceptan su participación *ex post* en caso de surgir vacantes. En segundo lugar, precisamente porque su razón de ser es la cobertura de vacantes, constituyen un mecanismo de aseguramiento para los gestores del programa contra dos tipos de contingencia: por un lado, «errores» en los mecanismos de selección de participantes, en el sentido de que algunos de los titulares abandonan el programa por falta de interés o motivación no detectadas correctamente; y, por otro lado, abandonos por condiciones sobrevenidas, no anticipables por los gestores, tales como el hecho de encontrar trabajo, un traslado de la familia, etcétera. A las entidades les interesa disponer de suplentes como «mecanismo de aseguramiento» porque si no se cubrieran las vacantes, se resentiría el indicador de gestión que valora como negativo un porcentaje elevado de abandonos.

En cualquier caso, más allá del porqué de su existencia, la cuestión importante desde la perspectiva de la evaluación es que el hecho de que haya suplentes abre la puerta a estrategias de estimación del impacto muy interesantes. En particular, si restringimos ahora nuestro foco de atención exclusivamente a los suplentes, el mecanismo de selección de participantes entre los suplentes «se parece» al de un sorteo: si el suplente está inscrito en un curso donde surgen muchas (pocas) vacantes, su probabilidad de participar es más alta (baja); y dado que no está en manos de los individuos alterar el número de «boletos» que tienen para la rifa, las características de aquellos suplentes que acaban participando son idénticas a las de los que finalmente no participan (incluidos los atributos no observables).

Nuestra propuesta es aumentar aún más el atractivo evaluativo de la existencia de suplentes mediante un protocolo explícito del SOC que, a grandes rasgos, propusiera a los proveedores seguir los pasos relacionados a continuación:

1. Definir un porcentaje obligatorio de suplentes (por ejemplo, uno por cada tres plazas) para

cada convocatoria. En el caso de que la entidad no alcanzara este porcentaje, debería justificar el por qué.

2. La definición de suplente debería ser dicotómica (sí o no), sin ningún sistema de puntos que ordenase a los suplentes entre sí. Podrían utilizarse baremos para clasificar a los candidatos, pero una vez definido el umbral que marca el inicio de la condición de suplente, la puntuación del individuo deviene irrelevante.

3. La elección de los suplentes para convertirse en participantes debería basarse en un sorteo.

No obstante, si se opta por no querer restringir la capacidad de las entidades para «ordenar» a los suplentes y basar el criterio de repesca en estas ordenaciones, sería necesario exigir entonces que el SOC supiera cuáles son el baremo y las puntuaciones. Esto permitiría intentar estimar el impacto mediante la técnica de la regresión discontinua comentada en el apartado 3, si bien las limitaciones en este caso podrían ser importantes.

4.5 LA MEJORA DE LAS BASES DE DATOS

El alcance y la calidad de los datos disponibles son uno de los principales condicionantes a la hora de realizar una buena evaluación de impacto. En el caso de las PAE, tal y como ponen de manifiesto las diversas evaluaciones realizadas por Ivàlua (2009, 2011 y 2013), el punto de partida es mucho mejor que el que se observa en otros ámbitos de actuación pública en Cataluña. En particular, existen tres bases de datos que pueden utilizarse para estimar cuantitativamente el impacto laboral de las PAE financiadas por el SOC[7]:

1. Base de datos de integración del SOC (SIPAO). Identifica a los participantes de todas las PAE financiadas por el SOC, así como a cada una de las entidades proveedoras de programas (entes locales, empresas, entidades sin ánimo de lucro, etc.). Incluye la fecha de inicio y finalización de la participación en el programa por parte de los beneficiarios, así como otras variables que permiten «cualificar» la naturaleza de la participación (abandono, motivo del abandono, grado de aprovechamiento, etc.). Asimismo, cuando el programa incluye proyectos de naturaleza diversa, el SIPAO puede aportar información, por ejemplo, sobre el sector de actividad de la acción formativa o sobre el tipo de empresa en el que se desarrollan las prácticas.

2. Base de datos SICAS-SISPE. Es la base de datos general de los demandantes de empleo de Cataluña. Los registros históricos corresponden al último día laboral de cada mes, desde mayo de 2005. Contiene información sociodemográfica básica (sexo, año de nacimiento, edad, nacionalidad, población de residencia y nivel formativo), sobre las preferencias y disposición a trabajar (restricciones de jornada, disposición a trabajar fuera del municipio o

comarca, número de empleos diferentes solicitados y tipo de empleo solicitado), historia de desempleo (fecha de inicio del periodo de desempleo actual), historia laboral (declaración del tiempo total trabajado y declaración del sector del último empleo), así como otros atributos relevantes para el empleo (percepción de la prestación activa, conocimiento de idiomas, declaración de discapacidades, etc.). No contiene información sobre algunas características relevantes, como la estructura del hogar (hijos menores, edad del hijo más pequeño, monoparentalidad, etc.), la historia completa de desempleo (número y duración de los periodos anteriores de desempleo), historia laboral anterior a 2005 o el tiempo que resta hasta agotar la prestación o el subsidio de desempleo.

3. Base de datos de contratos laborales. Estructurada en archivos mensuales, contiene los contratos y prórrogas realizados en el mes de referencia. Ofrece información tanto de las características sociodemográficas básicas de la persona contratada como de la empresa contratante y del tipo de contrato.

4. Base de datos de afiliación a la Seguridad Social. Estructurada en archivos trimestrales, correspondientes al último día laborable de cada trimestre, contiene información sobre las altas en la Seguridad Social, incluidos el tipo de contrato (lo que permite identificar si es indefinido o temporal) y el DNI de la entidad contratante (que permite discriminar si es privada, pública o una entidad sin ánimo de lucro). Al igual que las otras bases de datos, los registros históricos están disponibles desde 2005. .

El trabajo a realizar con estas cuatro bases de datos para evaluar el impacto de una determinada PAE, ya sea el programa Suma't o los Planes de Empleo, es siempre el mismo y consiste en construir una sola base datos que contenga: 1) todos los desempleados de Cataluña que estaban registrados en las OTG en el momento de iniciarse el reclutamiento de participantes (SICAS); 2) una variable que distinga quiénes de estos desempleados participaron en el programa objeto de la evaluación, quiénes en otras PAE y quiénes en ninguna (SIPAO); 3) diversas variables relacionadas con la participación en el programa, tales como la duración, el abandono, el tipo de actividades, etc. (SIPAO); 4) las características sociodemográficas de todas las personas del punto 1 (SICAS); y, finalmente, 5) los *outcomes* laborales de todas las personas del punto 1, hayan participado o no en el programa, desde su fecha de finalización hasta el último trimestre disponible en los cortes trimestrales de afiliación (Seguridad Social), así como vía los ficheros de contratos. El NIF-NIE de los individuos es el campo que permite enlazar las cuatro bases de datos entre sí (SIPAO, SICAS, SS y contratos) para obtener la base de datos analítica que acabamos de describir.

El principal problema del procedimiento anterior es que ha sido necesario repetirlo ex profeso para cada una de las evaluaciones de impacto encargadas por el SOC tanto a Ivàlua como a otras instituciones. Así pues, una primera mejora en el ámbito de los sistemas de información consistiría en desarrollar una base de datos integrada para la evaluación de las PAE, a partir del enlace de SICAS, SIPAO, Seguridad Social y contratos en un primer momento, y de los

datos de Educación en una segunda fase. En esta línea llevan ya trabajando un tiempo el SOC y el Observatorio del Departamento de Empresa y Empleo; de hecho, actualmente ya están integrados en el Datawarehouse del Departamento los ficheros del ámbito laboral. Sería muy beneficioso para incrementar las condiciones de evaluabilidad de las PAE acelerar la fase actual del proyecto, que es la de integrar la información de políticas activas de empleo. Aun así, hay un importante obstáculo a superar: la gestión de las PAE en Cataluña no ha contado nunca con un sistema bueno y general (que abarque todas las PAE) de información para la gestión, y esto limita la disponibilidad y calidad de los datos. En este sentido, sería también muy importante acelerar los trabajos realizados al respecto contemplados en el Plan director TIC del SOC. Además, siguiendo la estela de otros países de nuestro entorno —la de Alemania, por ejemplo, (Kruppe *et al*., 2008)—, un proyecto de estas características debería permitir a la comunidad investigadora utilizar la base de datos a voluntad, estipulando las cautelas legales oportunas, puesto que de esta manera se maximizaría la generación de conocimiento y se preservarían las garantías de independencia en el análisis.

Por otra parte, al margen de si se acaba apostando o no por la generación de una base de datos integrada como la que acabamos de esbozar, habría otros aspectos relacionados con el alcance y la calidad de la información que también se deberían intentar mejorar:

- Incorporar más variables a SICAS que pueden resultar de interés a la hora de mejorar la comparabilidad entre participantes y no participantes, especialmente si se utiliza la técnica del *matching*, tales como, por ejemplo: el estado civil, el número de hijos y sus edades, el estado de salud, etc.

- Enlazar con otras bases de datos administrativas, para mejorar la información disponible sobre la percepción de prestaciones y subsidios de desempleo (INEM) o de la cuantía de los salarios percibidos (IRPF o bases de cotización).

- Posibilidad de realizar encuestas sobre muestras de desempleados con el fin de enriquecer la información administrativa disponible. En el caso de Alemania, por ejemplo, existe desde el año 2008 una encuesta realizada sobre una muestra de 20 000 parados, extraídos de la base de datos integrada a la que nos hemos referido previamente, que aporta información sobre atributos relevantes a la hora de evaluar el impacto de las PAE (Caliendo et al., 2010): actitudes, factores psicológicos, alcance de las redes sociales, intensidad en la búsqueda de trabajo, etc.

4.6 MÁS ALLÁ DEL IMPACTO: LA EVALUACIÓN ECONÓMICA DE LAS PAE

La evaluación de impacto nos permite averiguar la efectividad de las PAE, es decir, si las personas que participan en ellas mejoran (o no) su situación gracias al programa. Así pues, si ante una determinada problemática, como pueda ser la baja empleabilidad de los jóvenes

sin la ESO, hemos puesto en marcha diversas políticas que afrontan el problema mediante aproximaciones alternativas (Suma't, NCNO, PCPI, etc.), la evaluación de impacto nos permite valorar, como mínimo, si cada uno de estos programas resulta efectivo en lo que respecta a la mejora de la inserción laboral, al incremento del capital formativo, etc.

Sin embargo, dado que los recursos disponibles para políticas activas de empleo han sido, son y serán limitados, es preciso ir un paso más allá y averiguar si el presupuesto asignado a cada uno de los programas, sean o no para jóvenes, permiten obtener el máximo impacto social posible del dinero que el contribuyente ha puesto a disposición del SOC para llevar a cabo la PAE. Desde esta perspectiva —que es la propia de la evaluación económica—, la comparación entre los diversos programas no puede hacerse únicamente de acuerdo con la magnitud del impacto conseguido en cada caso, sino que hay que considerar el coste de los recursos dedicados. En particular, la mayor o menor eficiencia de los programas depende, por tanto, de la relación que exista entre, por un lado, los efectos positivos sobre los *outcomes* de interés (trabajo, salarios, etc.) y, por otro, la cantidad de recursos que ha habido que movilizar para conseguir dichos impactos. Por tanto, el concepto de eficiencia es más amplio que el de efectividad, quedando subsumido el último en el primero. En otras palabras, la efectividad es una condición necesaria, pero no suficiente, para la eficiencia.

El comentario anterior hace referencia al denominado análisis coste-efectividad (CE), que es un tipo particular de evaluación económica, y ha sido escogido para simplificar la exposición. La sencillez del método radica en que únicamente hay que expresar en términos monetarios los costes del programa, mientras que los impactos positivos pueden expresarse en unidades naturales (inserciones laborales, retornos al sistema educativo, etc.). En cambio, la evaluación económica utiliza también otras metodologías más exigentes, como el análisis coste-beneficio (ACB), que obligaría a monitorizar no solo los costes de las PAE, sino también sus impactos positivos (denominados beneficios). La gran ventaja del ACB es que no solo permite comparar entre sí políticas activas de empleo que tienen impactos sobre *outcomes* diferentes, algo que el CE no puede hacer, sino que además permite comparar la rentabilidad de las políticas activas de empleo con otros programas de impulso económico o, incluso, con políticas públicas de otros ámbitos de intervención (salud, educación, etc.)[8].

A fecha de hoy, la penetración de la evaluación económica ha sido muy débil en el ámbito de las PAE a nivel europeo, tal y como destaca Smith (2009) en su comparativa entre la UE y los Estados Unidos. No obstante, hay que comenzar a mejorar la información sobre los costes de las actuaciones que se contemplan, iniciando el desarrollo de una genuina contabilidad analítica de las PAE, ya que el próximo paso después de la exigencia de evaluar el impacto de las PAE —una vez consolidado— será el requisito de evaluar su eficiencia. Y si no lo hacemos por convicción —como sería sensato— tendremos que hacerlo por obligación.

Notas:

[7] *También es posible medir el impacto educativo de aquellos programas, como los PCPI, cuyo objetivo explícito es favorecer el retorno de los participantes al sistema educativo. En este caso, la información disponible procede de los registros administrativos del Departamento de Educación, los cuales permiten identificar, a nivel de individuo (NIF-NIE), diversos outcomes de interés: matriculación en GES, en Ciclos Formativos de Grado Medio, graduación de ESO, etc.*

[8] *Véase Raya y Moreno (2013) para una introducción a la evaluación económica de políticas públicas.*

5. CONCLUSIONES

La situación convulsa que atravesamos tanto económica como socialmente constituye, paradójicamente, una oportunidad propicia para la evaluación. La disminución de los recursos disponibles para la mayoría de políticas públicas, incluidas las PAE, hace aflorar espontáneamente las preguntas que se encuentran en la raíz de la evaluación de impacto: ¿cuáles son los efectos de los distintos programas? ¿Son unos más efectivos que otros? Además, en el caso particular de las PAE, la presión para profundizar en el conocimiento de la efectividad de los programas ha sido más grande, puesto que la financiación de una parte de las actuaciones en este ámbito procede del presupuesto comunitario.

Averiguar cuál ha sido la efectividad de una determinada PAE no es, sin embargo, una tarea sencilla. A este respecto, el principal propósito de la guía ha sido comunicar a una audiencia no especializada la esencia de la evaluación de impacto: por una parte, hemos establecido que medir el impacto de una intervención equivale a averiguar si la política ha causado alguna mejora en el *outcome* de interés, lo que exige determinar cuál habría sido el *outcome* de los participantes en ausencia del programa (contrafactual); por otra parte, hemos descrito las principales técnicas existentes a la hora de evaluar el impacto y cómo pueden utilizarse para evaluar las PAE del SOC. Además, hemos intentado ilustrar que existen análisis, tales como los indicadores de seguimiento o las autovaloraciones, que, a pesar de su popularidad y aunque proporcionan información valiosa sobre un programa, no permiten decir nada sobre cuál ha sido el impacto de una determinada política. En síntesis, hemos intentado acercar a gestores y políticos al corazón de una disciplina de investigación social aplicada como es la evaluación de impacto, sin abundar en explicaciones técnicas complejas.

La evaluación de impacto, a pesar de basarse en la aplicación de técnicas de análisis propias de las ciencias sociales, no es un ejercicio académico o, mejor dicho, no debería ser únicamente un ejercicio académico. La legitimidad de la evaluación emana de su capacidad potencial para mejorar la efectividad de las políticas y programas evaluados. Así pues, no basta con tener evaluaciones de impacto técnicamente robustas, realizadas con buenos datos y por investigadores y analistas competentes; si además se quiere que los resultados de las evaluaciones ayuden a mejorar el diseño de las PAE, hace falta que todos los implicados en su desarrollo, ya sean políticos o técnicos, hayan impulsado decididamente la realización de las evaluaciones, entiendan su contenido y valoren su aplicabilidad. Esta guía ha sido escrita, de hecho, con el propósito de conseguir esta complicidad activa hacia la evaluación por parte de los responsables del SOC.

Somos conscientes de que incrementar el «conocimiento evaluativo» es una condición necesaria, pero no suficiente, para que los resultados de las evaluaciones de impacto acaben mejorando el diseño de las PAE, ya que existen múltiples factores (políticos, sociales, etc.) que determinan la fisonomía de estos programas, más allá de la efectividad. De hecho, dado que

la legitimidad para desarrollar las PAE reside en los responsables políticos, el reto no pasa en absoluto por arrinconar la influencia de estos otros factores y sustituirlos por el «veredicto» de la evaluación; más bien al contrario, se trata de medir adecuadamente la efectividad de nuestras PAE y que sean los políticos y los gestores los que, legítimamente, tomen las decisiones que estimen oportunas, haciendo caso o no de la evidencia, pero disponiendo de información veraz sobre los impactos reales de los programas de los que son responsables.

6. PARA OBTENER MÁS INFORMACIÓN

Hay varios documentos recientes que aspiran, al igual que esta guía, a presentar las principales técnicas de evaluación de impacto a una audiencia no especializada:

- Fitzsimons, E.; Vera-Hernández, M. «A practitioner's guide to evaluating the impacts of labor market programs». *World Bank Employment Policy Primer* (December 2009). http://goo.gl/bzpGg

- Card, D.; Ibarrarán, P.; Villa, JM. «Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide ». *IZA DP* nº 6085 (October 2011). http://is.gd/JXuFL0

D'altra banda, si es busca aprofundir en els detalls tècnics dels diversos mètodes d'avaluació d'impacte, la següent llista ofereix algunes referències d'interès:

(1) Bernal, R.; Peña, X. *Guía práctica para la evaluación de impacto*. Bogotá: Publicaciones CEDE. Universidad de Los Andes, 2011. http://goo.gl/k6QjT

(2) Gertler, P. [et al.] *La evaluación de impacto en la práctica*. Washington DC: World Bank Training Series, 2011. http://goo.gl/uKfs1

(3) Imbens, G. W.; Wooldridge, J. «Recent Developments in the Econometrics of Program Evaluation». *Journal of Economic Literature* (2009), 47: 5-86.

(4) Burtless, G. «The case for randomized field trials in economic and policy research». *Journal of Economic Perspectives* (1995), 9: 63-84.

Las lecturas (1) y (2) son dos libros excelentes sobre el tema de la evaluación de impacto. Ambos utilizan muchos ejemplos y están muy bien documentados, e incluso contienen los datos y los códigos de programas informáticos para utilizar las técnicas descritas. La lectura (3) es una revisión de los últimos avances metodológicos y la (4) es una discusión muy interesante sobre la evaluación experimental de políticas públicas.

BIBLIOGRAFÍA

Blundell, R.; Dias, M. C.; Meghir, C. [et al.] (2004). "Evaluating the employment impact of a mandatory job search program". *Journal of the European Economic Association*, 2(4), 569–606.

Caliendo, M.; Falk, A.; Kaiser, L. C. [et al.] (2011). "The IZA Evaluation Dataset: towards evidence-based labor policy making". *International Journal of Manpower*, 32(7), 731–752.

Card, D.; Ibarraran, P.; Villa, J. (2011). "Building and evaluation component for active labour market programs: a practitioner's guide". *IZA*. Retrieved from http://goo.gl/yR1VM

Casado, D. (2012). "Per què no avaluem les polítiques públiques com els fàrmacs? Una aposta per l'experimentació social". *Avaluació per al Bon Govern*, (3). www.avaluació.cat

Graversen, B. K.; Van Ours, J. C. (2008). "Activating unemployed workers works; Experimental evidence from Denmark". *Economics Letters*, 100(2), 308–310.

Haynes, L. et alt. (2012). Test, Learn and Adapt. Developing Public Policy with Randomised Controlled Trials. Cabinet Office. Behavioural Insights Team. Retrieved from http://is.gd/U29XIl

Heckman, J. J.; Smith, J. A. (1998). "Evaluating the welfare state" (No. 6542). Cambridge: *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w6542

Hendra, R.; Riccio, J. A.; Dorsett, R. [et al.] (2011). Breaking the low-pay, no-pay cycle: Final evidence from the UK Employment Retention and Advancement (ERA) demonstration (Vol. 765). Department for Work and Pensions.

Kluve, J. (2010). "The effectiveness of European active labor market programs". *Labour Economics*, 17(6), 904–918.

Kruppe, T.; Müller, E.; Wichert, L. [et al.] (2008). "On the definition of employment and its implementation in register data: the case of Germany". *Schmollers Jahrbuch*, 128(3), 461–488.

Lázaro, B. (2012). "Les noves formes d'inversió social com a motor d'innovació de les polítiques públiques". *Avaluació per al Bon Govern*, (4). www.avaluació.cat

Raya, J. M.; Moreno, I. (2013). Guia d'introducció a l'avaluació econòmica (En premsa). Barcelona: Ivàlua.

Sanz, J. (2011). *La Metodologia qualitativa en l'avaluació de polítiques públiques* (No. 8). Barcelona: Ivàlua. Retrieved from http://goo.gl/SXos2

Schlotter, M.; Schwerdt, G.; Woessmann, L. (2010). *Methods for causal evaluation of education policies and practices: An econometric toolbox* (No. 4725). Berlin: IZA Institute for the Study of Labour.

Smith, J. (2004). *Evaluating local economic development policies: Theory and practice. In Evaluating local economic and employment development: How to assess what works among programmes and policies*. (OCDE., pp. 287–332). Paris.

Smith, J. (2009). *What can the ESF learn from US evaluations of active labour market programs? Presented at the Evaluation and Performance Management of Job Training Programs: What Can the European Social Fund Learn from the WIA Experience?*. Michigan: APPAM.

ANEXO. EJEMPLOS DE EVALUACIONES DE POLÍTICAS ACTIVAS DE EMPLEO

EJEMPLO 1. DISEÑO EXPERIMENTAL

Impacto de provisión pública frente a privada de un programa de orientación y búsqueda de empleo en Francia

- **Publicación:** Behaghel, L., B. Crépon and M. Gurgand (2012). Private and public provision of counseling to job-seekers: Evidence from a large controlled experiment. IZA Discussion Paper 6518.
- **Método:** Diseño experimental
- **Lugar:** 22 regiones administrativas de Francia
- **Periodo:** Enero 2007 - Marzo 2008
- **Fuente de datos:** Registros administrativos acompañados de una encuesta telefónica

Contexto y diseño de la política

Los programas de orientación y búsqueda de empleo han recibido mucha atención en los últimos años, sobre todo porque las evaluaciones de estos programas muestran que son efectivos, especialmente si se los compara con otras PAE más tradicionales, tales como la formación o el empleo subvencionado. Al inicio del programa, la tasa de desempleo era solamente del 8,4 %, pero una parte importante de estos desempleados (30 %) habían estado en paro durante un año o más.

En el año 2006, el fondo de prestaciones de desempleo contrató a proveedores privados para realizar una nueva intervención con la idea de ahorrar dinero a la Seguridad Social. En concreto, a cambio de un pago por cada desempleado, los proveedores tenían que orientarles y ayudarles intensivamente en la búsqueda de empleo. Adicionalmente, este servicio privado tenía que servir de competencia al, hasta entonces, monopolio público e incrementar así su eficiencia, que era muy cuestionada en aquel momento.

Los 11 proveedores privados (básicamente, agencias de trabajo temporal) fueron escogidos mediante un proceso de subasta en cada una de las 22 regiones participantes. El esquema de pago a los proveedores era igual en todas ellas: un 30 % al inicio, un 35 % si el desempleado tardaba menos de seis meses en encontrar un trabajo y un 35 % si lo mantenía durante más de seis meses. Sin embargo, fruto de las diferencias en las subastas, el intervalo de pago por desempleado se situó entre los 3 000 y los 4 000 euros en los casos más favorables, y entre los 900 y los 1 200 euros en los más desfavorables.

La subcontratación de servicios públicos a intermediarios privados puede generar problemas, sobre todo cuando la calidad o el alcance de las acciones no pueden observarse adecuadamente. En el caso de los servicios de orientación y búsqueda de trabajo, la heterogeneidad de los trabajadores hace que dimensiones como el esfuerzo o las habilidades de cada uno de los desempleados sean muy difíciles de observar. La estructura del contrato con los proveedores privados puede generar, por tanto, incentivos perversos que contrarresten el potencial de eficiencia de las empresas privadas: por una parte, haciendo *cream-skimming* con los desempleados, es decir, reclutando únicamente a aquellos con buenas posibilidades de encontrar trabajo; y, por otra, «aparcando» a los desempleados más difíciles de insertar, ofreciéndoles un mínimo de servicios.

Contexto y diseño de la evaluación

Durante 2006, el fondo de Seguridad Social aseguró que la experiencia con los agentes privados había sido muy exitosa. Desafortunadamente, esta afirmación se basaba en indicadores de empleo sin considerar la situación contrafactual, algo arriesgado dado el posible sesgo de selección derivado del hecho de que solamente un 30 % de los desempleados asignados a los proveedores privados acabaron participando. Así las cosas, con el propósito de obtener una medida creíble del impacto del programa, se llevó a cabo una evaluación experimental con asignación aleatoria de la participación.

La evaluación quería contestar a dos preguntas. En primer lugar, buscaba conocer si el mecanismo de orientación intensiva (cada trabajador social tiene un máximo de 40 desempleados asignados y como mínimo un contacto semanal, por teléfono o e-mail, y uno mensual, cara a cara, con el desempleado) era efectivo respecto al itinerario anterior (120 desempleados por trabajador con un único contacto mensual). La otra pregunta que interesaba responder era si el servicio privado era realmente más efectivo que el servicio público.

Dado que los proveedores privados solamente gestionaban un subconjunto de los desempleados de la agencia pública, concretamente los que llevaban menos de tres meses en el paro y que percibían una prestación de desempleo, la evaluación se concentró únicamente en este grupo. Así, cuando un desempleado de estas características asistía por primera vez a su correspondiente oficina de empleo, una vez analizado su perfil, se cargaba en una extranet que asignaba aleatoriamente al desempleado a uno de los tres grupos siguientes: prestador privado (grupo de tratamiento 1), prestador público con un itinerario intensivo (grupo de tratamiento 2) o el servicio público tradicional (grupo de tratamiento 3). Así pues, el diseño de la evaluación no solo permite analizar el efecto de la intervención intensiva (2 frente a 3), sino también comparar la eficiencia de los agentes privados con la de los agentes del servicio público a la hora de atender perfiles de desempleados similares (1 frente a 2).

Resultados

En primer lugar, según los resultados de la evaluación, la orientación intensiva en la búsqueda de trabajo incrementa la transición hacia el empleo tanto cuando se lleva a cabo en el servicio público como cuando la realizan proveedores privados. En concreto, la tasa de salida del paro se incrementa de 4 a 9 puntos porcentuales respecto a la intervención tradicional.

En segundo lugar, los impactos son consistentemente más elevados para los participantes que fueron atendidos por proveedores públicos. El análisis cualitativo explica este resultado por el hecho de que los proveedores privados «aparcan» a los desempleados cuya inserción es más difícil, un hecho que los autores relacionan con la elevada cuantía del pago inicial (de 900 a 1 200 euros). Una evidencia de este fenómeno, constatada a partir de la encuesta telefónica, es que los proveedores privados tratan de manera diferente a las personas según su perfil. Por ejemplo, los proveedores privados tienen más contactos con unos participantes que con otros, mientras que en la agencia pública los contactos y las entrevistas son más uniformes.

EJEMPLO 2. REGRESIÓN DISCONTINUA

La mejora de la inserción laboral de los jóvenes desempleados en el Reino Unido

- **Publicació:** De Giorgi, Giacomo (2005). *The New Deal for Young People Five Years On*. "Fiscal Studies", vol 26 (3) 371-383.
- **Método:** Regresión discontinua
- **Lugar:** Reino Unido
- **Periodo:** Abril 1998 - Diciembre 2003
- **Fuentes de datos:** Registros administrativos del programa (New Deal Evaluation Database, NDED) combinados con el registro de desempleados del Reino Unido (JUVOS)

Contexto y diseño de la política

Durante la década pasada, se comenzó a implementar en el Reino Unido una serie de políticas destinadas a incrementar la empleabilidad y habilidades de los beneficiarios de programas sociales, al tiempo que se ligaba la percepción de las prestaciones a la participación en políticas de activación. Uno de estos programas es el New Deal for Young People, de participación obligatoria, destinado a jóvenes de entre 18 y 24 años que hayan estado en paro durante seis meses como mínimo y que reciban una prestación de desempleo.

Este programa consta de dos etapas. Durante la primera, que dura cuatro meses, se acompaña

al desempleado en la búsqueda intensiva de trabajo, formación en habilidades básicas y confección de un CV. Esta etapa también incluye reuniones con el mentor una vez cada dos semanas en la que los jóvenes deben informar sobre qué actividades han estado haciendo para encontrar trabajo. En caso de que esta etapa no sea exitosa, es decir, que el desempleado no encuentre trabajo, se activa la segunda etapa, durante la cual, de común acuerdo con el trabajador social, pueden escogerse distintas opciones. La primera de ellas pasa por colocar al joven en un trabajo subvencionado durante seis meses. Una segunda opción, más adecuada para jóvenes sin formación básica, son los cursos de formación de 12 meses de duración, durante los cuales el joven continúa cobrando la prestación de desempleo. La tercera opción es un trabajo de voluntario donde el participante recibe durante seis meses la prestación de desempleo y un complemento. Finalmente, la cuarta opción es un trabajo en el sector público pensado para los perfiles más difíciles de colocar. Adicionalmente, acabada la segunda etapa, el programa puede ampliarse durante tres meses más con unas acciones muy parecidas a las de la primera etapa.

Contexto y diseño de la evaluación

El programa comenzó aplicarse en el año 1998 y desde aquel momento hasta diciembre de 2003 habían participado más de un millón de jóvenes, el 75 % de los cuales eran hombres. El piloto de este programa ya se había evaluado previamente con unos resultados positivos: el programa incrementaba la probabilidad de encontrar trabajo en 10 puntos porcentuales. La evaluación que ahora comentamos, en cambio, intenta mirar el efecto del programa cuando ya estaba completamente implementado, al objeto de comprobar si los buenos resultados de la fase piloto se mantenían o no.

En particular, de cara a identificar correctamente el impacto del programa, se utilizó una discontinuidad existente en su diseño: únicamente los jóvenes menores de 25 años podían participar en él. Así, bajo el supuesto de que los individuos que se encuentran alrededor de los 25 años de edad tienen una evolución similar, esta discontinuidad nos permite medir (localmente) el efecto de la política, ya que las personas que, por ejemplo, tienen 24 años y 11 meses estarán obligadas a participar, mientras que las que tienen 25 años y un mes no lo harán. Puede asumirse, no obstante, que este umbral de edad es básicamente arbitrario y que, por tanto, parece plausible pensar que no tiene ninguna otra consecuencia en el mercado laboral. De hecho, cuando se comparan otras características individuales como, por ejemplo, la experiencia laboral, el nivel educativo, la nacionalidad, etc., los dos grupos son prácticamente idénticos.

Resultados

La intervención se define como el total de etapas del programa y no como cada una de las diferentes opciones de la segunda etapa. El *outcome* de interés para evaluar el impacto es la

probabilidad de encontrar trabajo durante los 18 meses posteriores al inicio del programa (es decir, una vez agotados los seis meses de paro).

Los resultados del programa son positivos según la evaluación. En particular, como consecuencia de participar en el programa, los jóvenes de menos de 25 años tienen una probabilidad de encontrar trabajo durante los 18 meses posteriores seis puntos porcentuales más elevada que otros jóvenes con las mismas características, pero justo por encima de los 25 años. Así pues, una vez implementado completamente, el programa sigue teniendo los efectos positivos que ya había detectado la evaluación llevada a cabo durante la fase piloto.

EJEMPLO 3. MATCHING

Los efectos en Alemania de las políticas activas de empleo para jóvenes desempleados

- **Publicación:** Caliendo, M., S. Künn and R. Schmidl (2011). *Fighting youth unemployment: The effects of active labor market policies*. IZA Discussion Paper 6222.
- **Método:** *Matching*
- **Lugar:** Alemania
- **Periodo:** 2002- 2008
- **Fuentes de datos:** Registros administrativos del IAB

Contexto y diseño de la política

Al entrar en el mercado laboral, los jóvenes deben sortear demasiados obstáculos para encontrar trabajo, una situación que da lugar a una probabilidad de permanecer en paro significativamente más elevada que la de la media de la población. En este contexto, debido a las consecuencias negativas que tiene el que una parte considerable de la población esté en paro durante largos periodos de tiempo (una reducción permanente de la probabilidad de encontrar trabajo, un incremento de los delitos, una reducción del capital social, etc.), los gobiernos de la mayoría de países europeos han realizado un gasto considerable en políticas activas de empleo para jóvenes, con el objetivo de integrarlos en el mercado laboral y/o favorecer su retorno al sistema educativo.

En Alemania, en el periodo entre 2000 y 2010, 1,4 millones de jóvenes participaron cada año en políticas de activación, tanto en las destinadas a mejorar la oferta laboral (por ejemplo, asistencia en la búsqueda de trabajo, formación ocupacional de corta y media duración, medidas destinadas a promover la participación en formación ocupacional y programas de creación de trabajo), como en las destinadas a incentivar la demanda de empleo (por ejemplo, políticas de

subsidio a la contratación y programas de becarios). El objetivo de esta evaluación era conocer qué impacto había tenido a corto, medio y largo plazo el conjunto de políticas activas de empleo para jóvenes, tanto en lo relativo a la integración en el mercado laboral como al retorno a la educación formal.

Contexto y diseño de la evaluación

La evaluación del impacto del programa se hizo para jóvenes de menos de 25 años que entraron en las listas del paro durante el año 2002. Los participantes son todos aquellos que participaron en alguna de las políticas de activación; los no participantes, que sirven como potenciales controles, son los desempleados que durante sus primeros 12 meses en paro no habían participado todavía en una PAE. Finalmente, para decidir quiénes de estos desempleados actuaban como controles se utilizó la metodología del *matching* o emparejamiento. En concreto, las variables que se consideraron en el *matching* fueron las siguientes:

A) características sociodemográficas: sexo, edad, estado civil, situación familiar y nacionalidad;

B) nivel educativo y estado de salud;

C) información sobre la última actividad laboral: experiencia laboral previa, tipo de contrato en la última actividad laboral antes del paro, grupo ocupacional del último empleo, ingreso diario en el último empleo

D) historial laboral de los últimos tres años: meses en paro, meses trabajando, meses en una PAE, meses de inactividad, meses con un contrato de tiempo parcial

E) información sobre la condición actual de desempleado e información sobre el trabajador social asignado: meses restantes de prestación de desempleo, meses desde el último contacto con la oficina de empleo, información sobre preferencias ocupacionales, número de ofertas propuestas por el trabajador social

F) características regionales: tasa de desempleo y crecimiento del PIB en el último año

El total de participantes evaluados fue de 12 400, el de no participantes de 38 639 y el periodo de evaluación abarcó 60 meses a partir del inicio de la participación en cada uno de los programas.

Resultados

En términos de mejora de las posibilidades de encontrar trabajo por parte de los jóvenes desempleados, la perspectiva del conjunto de políticas de activación es positiva, ya que la

evaluación encuentra evidencia de un efecto persistente y estable en casi todos los tipos de medidas consideradas. Los programas que tienen un efecto más importante a largo plazo sobre la participación en el mercado laboral son los subsidios salariales (10 a 20 pp), seguidos de la asistencia en la búsqueda de empleo y la formación de corta y media duración (5 a 10 pp). Por el contrario, los programas de creación de empleo público reducen la probabilidad de encontrar trabajo a medio plazo. Finalmente, en lo que respecta al retorno al sistema educativo y, más concretamente, a los itinerarios de formación profesional, los programas de becarios parecen tener un impacto positivo, sobre todo a medio plazo.

En general, aun siendo efectivas, la evidencia muestra que las políticas más efectivas de todas están diseñadas para personas con un nivel educativo significativamente superior al de la media de los desempleados. En este sentido, hay que subrayar que únicamente las políticas de incentivo a la demanda (subsidios salariales) son efectivas para los jóvenes con un nivel educativo más bajo. Sin embargo, según señalan los propios autores, este colectivo es precisamente el más numeroso en términos de desempleados y, por tanto, sobre el que debería centrarse principalmente el diseño de las PAE para jóvenes.

EJEMPLO 4. MODELO DE DOBLES DIFERENCIAS

¿Reducen los subsidios el desempleo entre las personas de más edad?

- **Publicación:** Boockmann, B., T. Zwick and A. Ammermüller (2012) "Do hiring subsidies reduce unemployment among older workers? Evidence from natural experiments". *Journal of the European Economic Association*, vol 10 (4) 735-764.
- **Método:** Dobles diferencias
- **Lugar:** Alemania
- **Periodo:** 2001 - 2004
- **Fuentes de datos:** Registros administrativos del IAB

Contexto y diseño de la política

Los trabajadores de más edad suelen tener muchas dificultades para encontrar trabajo en caso de quedarse en paro. En el año 1998, con la intención de mitigar este problema, el Gobierno alemán introdujo un subsidio a la contratación de personas de 50 años o más, siempre y cuando llevasen como mínimo seis meses en paro durante los 12 meses previos a su contratación. Esta subvención representaba el equivalente a un 50 % de los costes laborales homogeneizados durante un periodo de 24 meses. A partir del año 2002, el Gobierno extendió esta subvención a todas las personas desempleadas de más de 50 años.

Contexto y diseño de la evaluación

El cambio regulatorio permite utilizar la metodología de dobles diferencias para estimar el efecto de los subsidios sobre la transición al empleo durante los primeros 180 días de paro. En concreto, la reforma generó dos grupos de características muy similares: uno quedó afectado por el cambio regulatorio (personas con 50 años o más) y el otro no (personas con menos de 50 años). En particular, teniendo en cuenta que las personas que llevaban en paro más de seis meses ya tenían derecho a ser contratados con la subvención, los dos grupos de interés son los desempleados de «50 a 50 más seis meses» y los de «49 a 49 y seis meses». Así pues, asumiendo que estos dos grupos se hubieran comportado de la misma manera en ausencia del cambio regulatorio, la comparación de los dos grupos antes y después permite saber si la transición desde el desempleo se redujo a partir de la introducción de la subvención. Dado que el cambio se produjo en el año 2002, el primer experimento es analizado a partir de los desempleados entre el 1 de abril y el 30 de junio de 2001 y de 2002.

Resultados

Según el estudio, la probabilidad de salir del paro y volver al empleo no se ve afectada por la disponibilidad de subsidios, excepto para las mujeres de Alemania del Este. Esta falta de efectividad es debida principalmente a los «pesos muertos» del programa: los incrementos en la contratación subvencionada vienen acompañados de la reducción de las contrataciones no subvencionadas.

ivàlua
Institut Català d'Avaluació de Polítiques Públiques
Pau Claris, 108
4t 1a
08009 Barcelona
T. +34 93 554 53 00
info@ivalua.cat
www.ivalua.cat

Exemplar impreso: 12€
Versión electrónica: www.ivalua.cat