Avaluació per al Bon Govern
Número 3. Octubre 2012
ivàlua | Institut Català d'Avaluació de Polítiques Públiques

¿Per què no avaluem les polítiques públiques com els fàrmacs? Una aposta per l'experimentació social - David Casado

David Casado és doctor en Economia per la Universitat Pompeu Fabra. Com a analista d'Ivàlua, on s'incopora el 2009, ha participat en l'elaboració de diverses de les guies metodològiques, ha estat formador en les diverses edicions del Cicle de Formació en Avaluació de Polítiques Públiques i ha participat en diverses de les avaluacions realitzades per aquesta institució.

Introducció

La crisi econòmica ha suscitat un interès renovat per l'avaluació de polítiques públiques o, si més no, pels conceptes sobre els quals versa. Així, ja sigui en l'àmbit estatal, autonòmic o local, els responsables polítics, independentment del seu color, insisteixen en la necessitat de «millorar l'efectivitat» de certs programes o «guanyar en eficiència» en la prestació dels serveis públics. Aquesta insistència ens sembla positiva, tot i que dubtem que aquest èmfasi es mantingui quan vinguin temps millors, fet que seria un error: preguntar-se si una determinada política activa augmenta la inserció laboral dels aturats (*efectivitat*) o si, per comparació a d'altres intervencions, el cost per aturat inserit d'aquesta política és més o menys favorable (*eficiència*) són qüestions que l'Administració s'hauria de plantejar en tot moment, sigui quina sigui la seva situació econòmica i l'estat de les finances públiques.

Això no obstant, a pesar d'aquest apogeu de l'avaluació en el pla discursiu, el cert és que el nostre país segueix estant a la cua dels països desenvolupats pel que fa a l'avaluació de polítiques públiques (Viñas, 2009). I, el que és fins i tot més preocupant, quan a l'empara d'aquest interès renovat per l'avaluació es discuteix sobre l'impacte d'una o altra política, els resultats que s'invoquen acostumen a basar-se, en el millor dels casos, en lectures errònies de les dades disponibles, quan no en anècdotes d'impossible generalització o en apriorismes ideològics sense cap base empírica.

Deixant de banda l'evidència anecdòtica i els apriorismes, dels quals el lector té segurament diversos exemples, centrem-nos en la qüestió de la mala interpretació de les dades. El Pla Prepara, que va acaparar bona part de les portades durant el passat mes d'agost, és un exemple revelador en aquest sentit[1]. Entre altres motius, la necessitat de reformar el programa es justificava pel fet que només el 6% dels beneficiaris assolia la reinserció laboral. Significa això que el programa no és efectiu? Bé,

depèn de quin sigui el percentatge d'aquests beneficiaris que, a manca del Pla Prepara, s'haguessin reinserit: si la resposta fos que el 6%, podríem concloure que en termes d'inserció laboral el programa resulta inefectiu; però i si aquest percentatge fos del 0%? Llavors el Pla Prepara seria el responsable que 6 de cada 100 beneficiaris trobés feina en lloc de seguir a l'atur com la resta. Però como saber quants dels beneficiaris del Pla Prepara haurien trobat feina si no haguessin participat en el programa?

Donar resposta a la pregunta anterior constitueix el gran repte a què s'enfronten aquells que es dediquen a l'avaluació d'impacte. Des d'aquesta perspectiva, l'impacte d'una intervenció o programa és la diferència entre allò que realment esdevé als participants i l'anomenat contrafactual: és a dir, allò que els hagués succeït si no haguessin participat. Es tracta d'un gran repte perquè, òbviament, no és possible que els mateixos subjectes participin i no participin de manera simultània en un determinat programa. Per aquest motiu, com va posar de manifest Marcos Vera en un número anterior d'aquesta revista, els avaluadors intenten aproximar-se a la mesura d'aquest contrafactual mitjançant la utilització de tècniques diverses que, tanmateix, comparteixen una característica comuna: comparar l'evolució dels *outcomes* d'interès, com per exemple la inserció laboral, entre els participants en el programa i un altre grup de persones que, malgrat no haver-hi participat, són molt similars a aquelles que sí ho han fet. Això no obstant, entre els diferents dissenys avaluatius existents, n'hi ha un que sobresurt per sobre de la resta: els experiments socials (d'ara endavant, «ES»).

A les pàgines següents intentarem explicar en què consisteix un ES i per què, tot i la seva senzillesa, constitueix el mètode més robust per estimar l'impacte d'una política pública. Així mateix, descriurem quin abast tenen en el món aquest tipus d'avaluacions, les principals crítiques que esgrimeixen els seus detractors i, finalment, quin pot ser el seu futur en el nostre país, on fins ara el seu nivell de penetració ha estat nul.

Què són els ES i per què no tenen rival a l'hora de mesurar impactes?

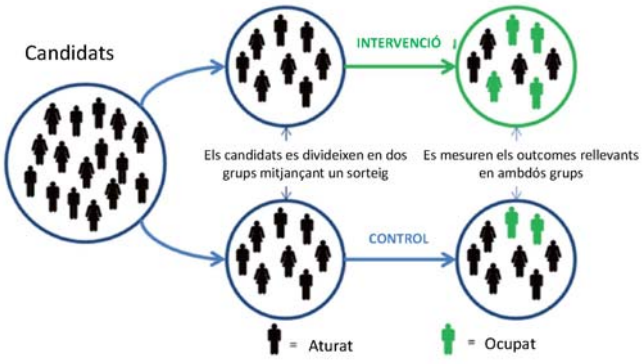
Suposem que el Pla Prepara no fos un programa laboral, sinó un nou fàrmac contra un càncer incurable, i que el 6% esmentat anteriorment es referís no a la taxa d'inserció laboral, sinó al percentatge de persones tractades que sobreviuen al cap d'un any. Com respondrien els metges a la pregunta sobre si el nou tractament resulta o no efectiu? En prin-

cipi, atès que fa vàries dècades que la professió mèdica basa els seus judicis sobre efectivitat en la realització d'assajos clínics, és probable que també en aquest cas haguessin procedit de la mateixa manera. Així, lluny de pronunciar-se sobre si una taxa de supervivència del 6% és baixa o alta, s'haurien preguntat quin percentatge de pacients hagués sobreviscut si no haguessin estat tractats amb el nou fàrmac. En concret, després de reclutar, per exemple, 1.000 pacients disposats a participar en l'assaig, haguessin subministrat el fàrmac a 500 d'ells escollits a l'atzar, mentre que als 500 restants els haguessin administrat un placebo. Un any després, haguessin comparat la taxa de supervivència del grup de tractament (6%) amb la del grup de control (*x*%) i declarat, senzillament, que el fàrmac resulta efectiu o no en funció que *x* fos o no menor que un 6% [2].

En essència, un experiment social és idèntic a un assaig clínic, amb l'única diferència que el «tractament» no és una intervenció sanitària, sinó un programa o política que aspira a produir canvis sobre determinats *outcomes* que la societat valora (incrementar la inserció laboral dels aturats, reduir el fracàs escolar, etc.).

Un exemple pot resultar útil per il·lustrar el funcionament d'un experiment social. Suposem una hipotètica política activa d'ocupació dirigida a joves aturats, d'entre 16 i 24 anys, que no tinguin el graduat en ESO. El programa en qüestió, que podríem anomenar *Segona Oportunitat,* consistiria en un número determinat de sessions formatives seguides per unes pràctiques de 6 mesos remunerades. Es podria dur a terme una avaluació experimental d'aquest programa, obviant de moment múltiples detalls als quals farem referència més endavant, de la següent manera: 1) donar instruccions a les oficines d'ocupació perquè identifiquessin potencials beneficiaris del programa, amb l'objectiu d'assolir una xifra de 1.000 candidats; 2) mitjançant un procediment aleatori, i previ consentiment dels candidats, aleatoritzaríem la participació en *Segona Oportunitat*: 500 joves el rebrien i 500 no; i 3) transcorregut un cert temps després de la finalització del programa, compararíem els *outcomes* rellevants, com el grau d'inserció laboral, entre el grup de tractament i de control. La figura 1 mostra gràficament l'essència d'una avaluació experimental del programa *Segona Oportunitat*.

Figura 1. Avaluació experimental del programa Segona Oportunitat



Font: Adaptat de Haynes et al. (2012: pàg. 9).

Ara bé, per què l'aleatorització, ja sigui de pacients en un assaig clínic o de joves aturats en el nostre exemple, permet avaluar l'impacte d'un tractament o programa d'una forma més vàlida que altres tècniques?

Gràcies a l'aleatorització, un experiment aconsegueix que el grup de tractament i el de control estiguin «equilibrats» en tots aquells atributs personals que poden influir sobre l'*outcome* d'interès, com poden ser, en el cas de *Segona Oportunitat*, la motivació, l'experiència laboral prèvia o el fet de ser o no immigrant [3]. D'aquesta manera, quan una vegada finalitzat el programa comparem els *outcomes* entre tots dos grups per inferir l'impacte, podem descartar que el resultat obtingut sigui la conseqüència de que ambdós grups són diferents. D'altra banda, atès que els dos grups estan exposats als mateixos «factors de context» mentre dura el programa, com podria ser per exemple una millora del mercat laboral en el cas de *Segona Oportunitat*, també podem descartar que aquests factors siguin els responsables de les diferències posttractament en els *outcomes*. En resum, si detectem aquestes diferències en els *outcomes* entre ambdós grups, els podrem atribuir a l'únic tret que els diferencia: és a dir, haver participat o no en el programa. L'experiment social ens proporciona, per tant, una estimació vàlida de l'impacte del programa.

Existeixen dissenys avaluatius no experimentals que també utilitzen grups de comparació, com poden ser el matching o el model de dobles diferències, fet que els permet tenir en compte la influència d'aquests factors contextuals. Això no obstant, en la mesura en què el procés de participació en el programa no és aleatori, la condició que ambdós grups tinguin característiques molt similars no està garantida.

Tornem a l'exemple de *Segona Oportunitat* per intentar il·lustrar aquest punt. Suposem que no es tracta d'un programa nou, sinó que porta alguns anys en funcionament, i que ens encarreguen avaluar l'impacte que ha tingut

sobre els joves que hi han participat. A més, com de fet acostuma a ser habitual, suposem també que la participació en el programa no ve determinada per un procés d'assignació aleatori, sinó que són els treballadors de les oficines d'ocupació qui seleccionen els candidats a participar i aquests últims decideixen, lliurament, participar-hi o no. En principi, per a totes les variables registrades en les bases de dades del Servei d'Ocupació, com són el sexe, l'edat, el nivell formatiu i moltes altres característiques dels individus, és possible identificar un grup de joves aturats que no hagi participat en el programa i que sigui semblant al grup d'aquells que sí hi han participat.

Ara bé, què succeeix amb totes aquelles variables sobre les quals no disposem d'informació, *inobservables* en termes tècnics, que poden haver influït sobre les decisions tant dels treballadors de les oficines d'ocupació com dels mateixos joves (motivació, implicació dels pares, renda familiar, etc.), i que clarament afecten les possibilitats d'inserció laboral posterior? Constitueixen un problema per a tots els dissenys no experimentals, ja que la seva influència sobre els *outcomes* és indistingible de l'impacte que realment té el programa, originant un error en el càlcul d'aquest impacte conegut com a «biaix de selecció». El gran avantatge dels ES és que aquest biaix queda eliminat, ja que el procés de participació ve definit per un procés d'assignació totalment aleatori i, com s'ha esmentat anteriorment, el grup de tractament i el de control estan equilibrats en tots els atributs que poden influir sobre els *outcomes* d'interès (fins i tot encara que siguin inobservables!). Pel que fa a la resta de dissenys no experimentals, per molt ben fets que estiguin, sempre existeix una ombra de dubte sobre fins a quin punt l'investigador ha estat capaç d'eliminar completament l'amenaça d'aquest biaix o, en altres paraules, descartar la possibilitat que l'impacte estimat de la política no es degui al fet que els participants són diferents de les persones amb qui se'ls ha comparat.

Què s'entén exactament per aleatorització?

L'aleatorització de la participació constitueix la pedra angular d'un ES i, perquè una avaluació pugui ser considerada experimental, la seva existència és un requisit indispensable. L'aleatorització que caracteritza un ES no s'ha de confondre amb el mostratge aleatori que s'ha d'exigir a una enquesta, ja sigui de salut o de població activa, perquè els resultats obtinguts siguin representatius de la població. D'una banda, mentre que el que ha de

ser aleatori en una enquesta és la selecció dels subjectes a entrevistar, en un experiment social s'exigeix que, entre els candidats a participar en el programa, l'elecció d'aquells qui hi acaben participant i d'aquells qui no es dugui a terme mitjançant un procediment aleatori. D'altra banda, l'aleatorització en un ES no busca la representativitat dels resultats, sinó permetre estimar sense biaixos l'impacte del programa avaluat, com ja hem comentat.

Això no vol dir, tanmateix, que un experiment social no pugui aleatoritzar també el procés de captació de candidats. Per exemple, si les 1.000 escoles de Catalunya estiguessin disposades a participar en un programa d'incentius a professors, però només hi hagués pressupost per aplicar l'esquema en 100 centres, podríem escollir 200 escoles a l'atzar d'entre les 1.000 (mostra de candidats) i, a continuació, assignar aleatòriament la participació en el programa a la meitat d'elles. La primera aleatorització conferiria «representativitat» als nostres resultats, en el sentit que es podrien considerar extrapolables a les 800 escoles «no experimentals», però és la segona aleatorització la que ens permet mesurar l'impacte del programa i l'única necessària per definir una avaluació com a experimental.

Una altra forma de dissenyar, implementar i redissenyar les polítiques públiques

Tot i que existeixen avaluacions experimentals de programes que porten anys funcionant, com per exemple la realitzada entre 1996 i 2003 del JobCorps dels Estats Units (Schochet et al., 2008), un programa dirigit a joves en risc d'exclusió social iniciat l'any 1960 i plenament consolidat, la majoria d'ES es plantegen sobre una política nova o sobre una variació d'un programa ja existent. De fet, a diferència de les avaluacions de tipus retrospectiu, en les quals es tracta d'estimar l'impacte que hagi pogut tenir una política una vegada finalitzada, els ES no només es dissenyen alhora que la política que es pretén avaluar, sinó que van de la mà durant tot el procés d'implementació. En el fons, com bé indica el suggeridor títol d'un recent informe sobre experiments socials, *Test, Learn and Adapt* (Haynes et al., 2012), es tracta d'una modalitat d'avaluació que obre les portes a una manera diferent de desenvolupar les polítiques públiques, si bé també planteja altres reptes importants als polítics i gestors que estiguin disposats a impulsar-los.

En primer lloc, acceptar avaluar experimentalment una nova política pública exigeix re-

conèixer, explícitament, que no se sap amb certesa si el programa en qüestió resultarà o no efectiu. Es tracta d'un exercici d'humilitat intel·lectual poc comú en el panorama polític actual, malgrat que existeixen innombrables exemples de polítiques els efectes de les quals han resultat ser nuls o fins i tot negatius: per exemple, en l'àmbit de la justícia, de les més de 80 avaluacions experimentals dutes a terme als EUA de programes de tot tipus, des de dispositius de reinserció de joves fins a modificacions en els tipus de sanció, ni més ni menys que el 81% van tenir resultats nuls o negatius (Farrington i Welsh, 2005). En altres ocasions, com revela l'avaluació d'un programa al Regne Unit que pretenia reduir la dependència dels beneficiaris de prestacions assistencials, ajudant els participants a conservar una feina quan finalment la trobaven, el programa va resultar ser efectiu per a qui menys s'esperava: així, si polítics, gestors i investigadors creien, abans de dur a terme l'avaluació, que el nou programa seria efectiu per a les famílies monoparentals però no per als aturats de llarga durada, els resultats foren exactament els contraris (Hendra et al., 2011) [4].

D'altra banda, a més d'humilitat intel·lectual, la naturalesa prospectiva dels ES obliga a polítics i gestors a explicitar, per endavant, quins són els *outcomes* sobre els quals el programa pretén incidir, així como a consensuar amb els avaluadors la manera com es quantificaran aquests *outcomes*. Així mateix, lluny de veure's limitats a emprar les dades que existeixin sobre el programa, com succeeix en les avaluacions de caràcter retrospectiu, els ES permeten als investigadors definir anticipadament quin tipus d'informació es requereix per dur a terme l'avaluació i, si és necessari, afegir alguna nova variable als registres administratius o complementar-ne la informació a través d'enquestes. Tots aquests elements redueixen les possibilitats de manipulació ex post dels resultats de l'avaluació, ja que els diferents elements han estat definits per endavant, per la qual cosa augmenta la credibilitat dels resultats obtinguts. A això caldria afegir-hi, com ja s'ha esmentat anteriorment, la superioritat dels experiments per mesurar l'impacte d'un programa de forma vàlida.

Finalment, de cara al desenvolupament de noves polítiques i al perfeccionament de les ja existents, els resultats de les avaluacions experimentals esdevenen un instrument socialment molt útil. La raó més òbvia és que ens permeten determinar, amb rigor, quines són les polítiques que funcionen i, per tant, acabar generalitzant únicament els progra-

mes pilot que han demostrat ser eficaços. Això no obstant, fins i tot si els experiments posen de manifest la inefectivitat d'una política, aquest resultat esdevé igualment valuós: ens permet comprendre per què la política no genera els efectes desitjats i, a continuació, proposar i comprovar experimentalment noves versions del programa que no ha funcionat.

Els experiments socials estan ja a l'altra banda dels Pirineus

El *New Jersey Income Maintenance Experiment*, dut a terme als EUA a finals dels anys 60 del segle passat, acostuma a considerar-se la primera avaluació experimental d'una política pública (Burtless y Hausman, 1978). El seu objectiu era analitzar en quina mesura els beneficiaris d'un programa de manteniment de rendes, similar als programes autonòmics de rendes mínimes, modificaven el seu comportament laboral davant diferents quanties de la prestació econòmica rebuda. A aquest primer experiment van seguir d'altres igualment famosos, com el *Rand Health Insurance Experiment*, realitzat a principis dels 80, també als EUA, amb l'objectiu de comprovar l'impacte sobre la salut i l'ús de serveis sanitaris de diferents configuracions de copagaments (Newhouse, 1993). En l'àmbit educatiu, destaca el conegut popularment com a Projecte STAR, dut a terme a Tennessee a finals dels 80, amb el propòsit d'analitzar experimentalment si la mida de les classes tenia algun impacte sobre el rendiment acadèmic dels alumnes a mitjà i llarg termini (Mosteller, 1995).

En qualsevol cas, a més dels experiments anteriors, la influència política i repercussió mediàtica dels quals fou especialment acusada, són múltiples els ES realitzats durant les darreres tres dècades als EUA en àmbits tan diversos com la justícia (Farrington y Welsh, 2005), l'educació (Antonio Cabrales, a «Nada es gratis», aquí o aquí) o els programes d'assistència social (Butler et al., 2012). De fet, segons l'inventari realitzat per Greenberg i Shroder (2004), el nombre d'experiments socials duts a terme als EUA en els àmbits assenyalats s'aproximaria als 300 durant el període 1980-2003. Cal dir que aquests experiments no sempre avaluen programes del sector públic, sinó que sovint es tracta d'avaluacions impulsades per organitzacions filantròpiques per comprovar l'efectivitat de les polítiques per elles finançades. Fins i tot en l'àmbit empresarial, com

detalla Manuel Bagües en aquesta altra entrada de «Nada es gratis», s'han produït alguns avenços substancials en l'aplicació d'avaluacions experimentals.

Un altre àmbit en què l'experimentació social ha registrat un avenç molt notable és el relacionat amb l'avaluació dels programes d'ajuda al desenvolupament, en especial els afavorits per organismes internacionals com el Banc Mundial o el Banc Interamericà de Desenvolupament. El llibre recent de Banerjee y Duflo (2011), que porta el suggeridor títol de Poor Economics, ofereix una panoràmica realment interessant del que ha donat de si l'experimentació social en l'àmbit dels programes d'ajuda al desenvolupament. També en alguns països d'Hispanoamèrica s'han dut a terme experiments d'una transcendència notable, entre els quals destaquen especialment els programes Progresa i Seguro Popular de Salud, ambdós realitzats a Mèxic: el primer per combatre la pobresa mitjançant transferències monetàries que s'havien de destinar a l'escolarització dels nens i a activitats de salut preventives (Schultz, 2004), i el segon per comprovar els efectes sobre la salut i l'ús de serveis sanitaris de l'ampliació de l'assegurament sanitari a famílies de baixos recursos (King et al., 2007).

A Europa, fins fa a penes una dècada, els ES eren una autèntica raresa. Això no obstant, durant aquests últims anys s'han començat a realitzar nombroses avaluacions experimentals, especialment en l'àmbit de les polítiques laborals, com l'Employment Retention and Advancement (ERA) demonstration duta a terme al Regne Unit (Hendra et al. 2011) o l'avaluació de diversos esquemes d'activació d'aturats realitzats a Dinamarca (Graversen i Van Ours, 2008), per posar només dos exemples. Una menció especial mereix el Fonds d'Expérimentation pour la Jeunesse, el qual inicià el seu camí l'any 2008 i que, a través d'una dotació de més de 200 milions d'euros per al període 2009-2014, intenta afavorir l'avaluació experimental de noves formes d'intervenció destinades a combatre el fracàs escolar i l'exclusió laboral dels joves francesos. Els programes avaluats fins ara, o en curs d'avaluació, són molt variats i inclouen des d'intervencions orientades a incrementar la implicació dels pares en el procés educatiu (aquí) fins a la provisió de tutories per prevenir l'abandó escolar (aquí), passant per la prestació de serveis d'orientació laboral a joves aturats (aquí).

Crítiques als experiments

En tractar-se del disseny més robust per avaluar l'impacte d'una política, i malgrat que el seu nombre ha crescut en els darrers anys, no deixa de sorprendre que no es duguin a terme molts més ES. En part, aquesta carència demostra l'efectivitat d'algunes de les crítiques vessades sobre els experiments, si bé el fonament lògic de moltes d'elles resulta quant menys qüestionable.

Un primer aspecte que se sol criticar dels experiments és que són cars. Sens dubte, aquesta apreciació es deu en part al pòsit que van deixar els primers ES duts a terme als EUA, com el New Jersey Income tax o el Rand Experiment esmentats anteriorment, els quals van implicar equips molt nombrosos, realització d'enquestes molt costoses, etc. Avui dia, com mostra a través de diversos exemples la Coalition for Evidence-Based Policy, la informatització dels registres administratius ha permès, en molts casos, esquivar la necessitat de realitzar costoses enquestes sense que això suposi perdre riquesa analística, ja que l'exhaustivitat i fiabilitat d'aquests registres resulta molt notable. Per exemple, com demostra l'experiment dut a terme per Fryer a Nova York analitzant l'impacte d'un programa d'incentius a professors, n'hi ha prou amb aleatoritzar les escoles i analitzar els *outcomes* a través dels registres administratius de l'autoritat educativa (proves estandarditzades, taxa d'abandó escolar, etc.).

En qualsevol cas, més enllà de les consideracions econòmiques, l'argument habitual que utilitzen els qui s'oposen als ES té un transfons ètic: resulta inadequat privar determinats individus (els del grup de control) dels beneficis que suposa una nova política utilitzant un mecanisme tan arbitrari com l'aleatorització. La rèplica per part d'aquells qui veuen en els ES una eina adequada d'avaluació se sustenta en tres consideracions.

La primera és que la presumpció que s'està privant alguns individus d'alguna cosa beneficiosa no hauria de tenir sentit si l'experiment està justificat, ja que és precisament l'absència de dades sobre l'efectivitat del programa el que justifica la seva avaluació. D'altra banda, són poques les ocasions en què pertànyer al grup de control implica no rebre cap tipus d'intervenció, sinó que més aviat el que es compara és la nova política respecte de «seguir com fins ara». Finalment, hi ha situacions força freqüents en què es pot considerar l'aleatorització com un

criteri d'assignació equitatiu, com per exemple quan la manca de recursos no permet atendre d'una sola vegada tota la població potencialment beneficiària de la política; de fet, quan es produeixen situacions d'aquest estil, un disseny experimental més acceptable que utilitzar una simple loteria entre individus és optar per un desplegament gradual aleatoritzat: s'aleatoritza el moment del temps en què diferents grups d'individus o territoris començaran a rebre el nou programa. Un enfocament d'aquestes característiques, por posar-ne un exemple, és el que es va dur a terme a Mèxic per avaluar experimentalment el programa Progresa esmentat anteriorment (Schultz, 2004).

Una altra crítica que s'acostuma a formular amb relació als experiments és que, tot i que permeten establir l'efectivitat d'una política, no resulten útils per comprendre per què la política funciona o no. Probablement la crítica fos encertada en el cas dels primers ES, molts d'ells de «caixa negra», en el sentit que s'aleatoritzava la participació i es mesuraven els resultats posttractament, sense parar-se a descriure el procés d'implementació del nou programa. Això no obstant, des de fa ja alguns anys, els experiments millor dissenyats acostumen a incloure una avaluació qualitativa i quantitativa de la implementació, la qual permet aventurar hipòtesis sobre quins components de la política poden haver tingut una major influència en els resultats d'impacte observats. Un bon exemple d'aquest tipus d'enfocament és el treball de Bloom et al. (2003) sobre els efectes dels programes d'activació dirigits als perceptors de prestacions assistencials, ja que els experiments realitzats no només van permetre establir l'efectivitat dels diferents programes, sinó també la major o menor influència que tingueren sobre aquesta efectivitat diferents components dels programes (grau de personalització de l'atenció, nombre de casos per treballador social, èmfasi a buscar feina per sobre de la formació, etc.).

Una última objecció que habitualment es fa als experiments és que acostumen a no tenir validesa externa o, en paraules menys tècniques, que els resultats que s'obtenen pel que fa a l'impacte d'una política, tot i ser vàlids respecte dels subjectes, moment i lloc en què es va dur a terme l'experiment, poden no ser extrapolables a contexts diferents. Aquells que es dediquen a l'experimentació social han intentat mitigar la manca de validesa externa per dues vies. En primer lloc, tot i que costa augmentar els recursos

necessaris, són habituals les avaluacions *multi-site*, en què el programa s'avalua aplicant-lo a llocs diversos (per exemple, escoles públiques i concertades, rurals i urbanes, etc.), amb l'objectiu d'analitzar fins a quin punt els resultats d'impacte varien en funció dels contextos. D'altra banda, quan el nombre de rèpliques experimentals d'un determinat tipus de programa és suficientment important, es pot dur a terme l'anomenada metanàlisi dels resultats obtinguts, és a dir, un exercici quantitatiu de síntesi que pretén establir si el programa resulta efectiu amb caràcter general, amb independència de les poblacions, llocs i moments en què s'aplica. En aquest sentit, esdevé crucial la tasca d'inventari duta a terme per organitzacions o iniciatives de diferent índole, com per exemple el What Works Clearinghouse, que avalua i sintetitza els resultats de tot tipus d'experiments duts a terme en l'àmbit educatiu (programes de lectura, de reforç escolar, d'atenció precoç, etc.).

Algunes precaucions

És possible que les crítiques als experiments no siguin adequades en molts casos, però no per això s'ha de pensar que dur a terme una avaluació experimental sigui una tasca senzilla. En primer lloc, des d'una perspectiva metodològica, cal ser conscient dels reptes que suposa realitzar una avaluació experimental i de les dificultats que poden sorgir. Algunes de les més importants són les següents [5]:

- La grandària dels grups. Una de les primeres qüestions a què ha de respondre un ES és quantes unitats, ja siguin alumnes, escoles o jutjats, han de formar part dels grups de control i tractament. Deixant de banda els detalls estadístics, direm simplement que hi ha dos factors que influeixen especialment sobre la qüestió de la grandària de les mostres: d'una banda, la magnitud de l'impacte que vulguem ser capaços de detectar (per exemple, augments de la inserció laboral d'un punt percentual) i, d'altra banda, el grau d'incertesa sobre la validesa dels nostres resultats que estiguem disposats a tolerar. De vegades, existeixen determinades intervencions en què n'hi ha prou amb aleatoritzar unes desenes d'unitats per extreure conclusions rellevants, mentre que en altres tipus de programes es requereixen diversos milers de subjectes per aconseguir el mateix. De totes maneres, és una qüestió que s'ha de tenir en compte des de l'inici de l'avaluació, ja que un experiment amb grandàries mostrals insuficients pot acabar sent inútil per detectar els efectes d'un programa, no perquè aquests no existeixin, sinó perquè

el nostre disseny no és capaç de detectar-los.

Risc de contaminació. Un problema amb el qual poden topar els ES és que, malgrat haver estat assignats aleatòriament als grups de tractament i control, alguns dels individus del primer grup acaben no rebent allò que el programa preveu (per exemple, perquè decideixen no assistir a les classes que Segona Oportunitat estipula) i/o alguns del grup de control acaben tenint-hi accés (per exemple, perquè els treballadors de les oficines d'ocupació sucumbeixen a les pressions d'alguns «no tractats»). El risc que es produeixi aquest tipus de situacions depèn, en gran mesura, de quina sigui la capacitat dels responsables de l'experiment per monitoritzar l'activitat dels gestors del programa i evitar situacions anòmales.

Externalitats. Qualsevol efecte indirecte sobre els outcomes del grup de control motivat per l'existència del tractament posa en dubte la validesa dels resultats generats per l'experiment. Una selecció precisa de les unitats a partir de les quals es realitzarà el procés d'aleatorització pot prevenir aquest tipus de biaix; per exemple, si estem interessats a mesurar l'impacte d'un programa escolar de salut alimentària sobre l'obesitat infantil, és evident que l'aleatorització no s'haurà de realitzar entre individus d'una mateixa escola (hi haurà processos d'imitació), sinó entre escoles que es trobin a una certa distància entre si.

En qualsevol cas, el caràcter prospectiu dels ES fa que les fases de planificació i disseny de l'avaluació siguin de cabdal importància. El risc de contaminació, l'existència d'externalitats o qualsevol altre factor que pugui esbiaixar els resultats de l'avaluació, hauran de ser anticipats i incorporats al disseny de l'experiment per intentar eliminar-los o, com a mínim, minimitzar-ne l'abast. En cas contrari, quan l'experiment ja està en marxa, resulta pràcticament impossible refer el disseny i la validesa dels resultats obtinguts pot quedar seriosament compromesa.

De totes maneres, més enllà de les qüestions tècniques que acabem d'esmentar, els veritables obstacles als quals s'enfronten els experiments solen ser sovint de caràcter politicoadministratiu. No hi ha constància de la multitud d'experiments que han estat descartats, moltes vegades al·legant els impediments esmentats anteriorment (ètica, costos, etc.), però de ben segur que es compten per centenars. A més, fins i tot quan s'acaba

duent a terme un experiment, res impedeix que un canvi de govern posi fi a l'experiment o que els responsables de la seva implementació en el terreny intentin sabotejar-lo [6]. Això no obstant, com apunten King et al. (2007), l'existència d'aquests i altres condicionants politicoadministratius no ha de ser vista com una anomalia, sinó que constitueix l'essència del terreny en què es desenvolupen les polítiques públiques i, per tant, els ES. Es tracta, per contra, de tenir-los en compte des del principi de l'experiment i, evidentment, descartar-ne l'aplicació si hi ha indicis clars que aquest no prosperarà.

El futur dels experiments socials a aquesta banda dels Pirineus

Els ES constitueixen un disseny avaluatiu d'una potència molt notable i el seu ús, quan és convenient i es realitza correctament, pot ajudar a desenvolupar polítiques públiques més efectives. Tanmateix, malgrat l'impuls que l'experimentació social ha viscut en les últimes dècades, sobretot als EUA però també en alguns països europeus, la seva utilització a Espanya ha estat fins ara inexistent. És cert que la seva implementació planteja reptes tècnics importants, i que sovint existeixen condicionants polítics a tenir en compte, però creiem que la situació de total absència d'ES en el nostre país no s'ha de mantenir durant més temps.

Estant com estem immersos en una crisi de les finances públiques molt preocupant, sembla obligatori impulsar el desenvolupament d'ES: permetrien quantificar l'efectivitat real de bona part de les polítiques públiques que estan ara sota sospita, i que s'eliminen o es mantenen sense cap evidència sobre la seva efectivitat, i també dels nous programes que sovint es proposen per millorar la situació. Així mateix, existeixen múltiples professionals amb capacitat tècnica suficient per dur a terme aquest tipus d'avaluacions, especialment en l'àmbit universitari. En darrer lloc, com posa de manifest l'article de Blanca Lázaro en aquesta mateixa revista, existeixen fórmules d'institucionalització de l'avaluació experimental que es podrien aplicar en el nostre context sense massa problemes (per exemple, el Fonds d'Expérimentation pour la Jeunesse esmentat anteriorment).

Ja s'ha recorregut una part del camí. En el fons, la majoria d'ES no són sinó programes pilot la generalització dels quals depèn, fonamentalment, de la capacitat que demostren de resultar eficaços. I els pilots, amb aquest mateix nom, no són aliens a la implementació de les nostres polítiques públiques: així,

per posar només un exemple, el programa Suma't del Servei d'Ocupació de Catalunya, que pretén incrementar la inserció laboral dels joves amb baixa formació, fou concebut com un projecte pilot. Això no obstant, lluny d'ésser vist com una oportunitat per comprovar l'efectivitat de la política abans de suggerir la seva generalització, els pilots en el nostre país són simplement «assajos» destinats a millorar-ne la implementació, atès que l'efectivitat es dóna per feta en la majoria del casos. Tanmateix, un cop s'ha determinat que el programa només s'aplicarà en alguns territoris, l'únic que separa els nostres «pilots» actuals dels ES és l'aleatorització. I aquest mecanisme d'assignació, com ja s'ha comentat, pot resultar fàcilment defensable quan, com acostuma a passar, la manca de pressupost impedeix aplicar el nou programa a tots els potencials beneficiaris.

Però més enllà de les qüestions anteriors, el gran repte és convèncer els responsables polítics i els gestors dels programes de les possibilitats que ofereix l'experimentació social. Aquesta sort de «conversió», per donar-li algun nom, exigeix dues condicions gens trivials: d'una banda, reconèixer que no se sap si una determinada intervenció resultarà o no efectiva i, d'altra banda, ser conscient que l'experimentació social és la manera més fiable, si es duu a terme correctament, d'avaluar si una cosa funciona o no. Som conscients que es tracta d'un canvi cultural de primera magnitud.

Ens fa l'efecte que el camí serà llarg i ple de dificultats, però no s'ha de pensar que no es pot recórrer, llevat que acceptem que hi ha quelcom intrínsec a nosaltres, alguna cosa genètica, que ens impedeix introduir l'avaluació experimental en les nostres polítiques públiques. Nosaltres creiem que no n'hi ha. De fet, amb aquest article, hem volgut aportar el nostre granet de sorra perquè abans o després s'arribi a invalidar, també en aquest camp, el vell tòpic que «*Spain is different*» i que, per fi, l'experimentació social acabi creuant els Pirineus.

Per saber més

- Descarrega't gratuïtament l'informe *Test, Learn and Adapt*, recentment publicat pel Cabinet Office del Regne Unit: http://is.gd/U29XIl
- No et perdis l'excel•lent curs gratuït sobre avaluació experimental impartit per Esther Duflo i altres membres del Poverty Action Lab (MIT): http://is.gd/yBeJLN
- Molt recomanable el web del What Works Clearinghouse: un magnífic repositori sobre intervencions en l'àmbit educatiu avaluades experimentalment: http://is.gd/_j3v2R4

[1]El Pla Prepara, per a aquells que han estat de vacances fora d'Espanya, és el programa que concedeix 400 euros als aturats de llarga durada que han esgotat la prestació o el subsidi d'atur, sempre i quan acceptin participar en accions formatives i/o d'orientació laboral.

[2]No sempre els tractaments sanitaris són sotmesos a l'escrutini d'avaluacions experimentals. En aquest sentit, un cas especialment dramàtic és el tractament amb esteroides administrat a persones que havien sofert un traumatisme craneal (Haynes et al., 2012). Aquesta pràctica, que s'havia utilitzat de manera rutinària durant dècades, fou sotmesa a un assaig clínic l'any 2004. Els resultats no només van demostrar tot el que tothom ja creia (és a dir, que era un tractament efectiu), sinó que van revelar que el grup de tractament estava experimentant un risc de mort major. De fet, l'assaig es va haver de suspendre per no seguir perjudicant els subjectes tractats.

[3] En termes tècnics aquest equilibri implica que, per a cadascuna d'aquestes característiques, no existeixen diferències estadísticament significatives entre la mitjana observada en un i altre grup. Consulti Duflo et al. (2007) per a una descripció formal dels fonaments estadístics dels experiments socials com a tècnica per mesurar impactes .

[4] Es poden trobar exemples de polítiques inefectives, o fins i tot perjudicials, en molts àmbits d'intervenció pública. Haynes et al. (2012) ofereixen exemples interessants en aquest sentit.

[5]Consulti Duflo et al. (2007) per a una anàlisi detallada en aquest sentit.

[6] En aquest sentit, és il·lustratiu un dels primers experiments duts a terme a Noruega en matèria de polítiques laborals (Torp et al., 1993). Gràcies a l'existència prèvia d'un excés de demanda generalitzat, amb més aturats que places disponibles, es va pensar que l'aleatorització seria factible. Tanmateix, després de posar en marxa l'experiment, els treballadors de les oficines d'ocupació, que havien de dur a terme el procés de selecció, van optar per identificar com a potencials candidats un nombre concret de persones que sempre coincidia amb el de places disponibles, i es va eliminar així la necessitat d'aleatoritzar la participació.

Bibliografía

Banerjee, A., & Duflo, E. (2011). ***Poor economics: a radical rethinking of the way to fight global poverty***. New York: PublicAffairs.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). **Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments**. *Journal of Policy Analysis and Management*, *22*(4), 551–575.

Burtless, G., & Hausman, J. A. (1978). **The effect of taxation on labor supply: Evaluating the Gary negative income tax experiment.** *The Journal of Political Economy*, *86*(6), 1103–1130.

Butler, D., Alson, J., Bloom, D., Deitch, V., Hill, A., Hsueh, J. A., Jacobs, E., et al. (2012). ***What Strategies Work for the Hard-to-Employ? Final Results of the Hard-to-Employ Demonstration and Evaluation Project and Selected Sites from the Employment Retention and Advancement Project*** (No. 2012-08). Office of Planning, Research and Evaluation (OPRE).

Duflo, E., Glennerster, R., & Kremer, M. (2007). **Using randomization in development economics research: A toolkit**. *Handbook of development economics*, *4*, 3895–3962.

Farrington, D. P., & Welsh, B. C. (2005). **Randomized experiments in criminology: What have we learned in the last two decades?** *Journal of Experimental Criminology*, *1*(1), 9–38.

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

Greenberg, D. H., & Shroder, M. (2004). ***The digest of social experiments***. Washington D.C: Urban Inst Press.

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

Hendra, R., Riccio, J. A., Dorsett, R., Greenberg, D. H., Knight, G., Phillips, J., Robins, P. K., 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.

Mosteller, F. (1995). **The Tennessee study of class size in the early school grades**. *The future of children*, *5*(2), 113–127.

Newhouse, J. P. (1993). ***Free for all?: lessons from the RAND health insurance experiment***. Cambridge: Harvard University Press.

Paul Schultz, T. (2004). **School subsidies for the poor: evaluating the Mexican Progresa poverty program.** *Journal of development Economics*, *74*(1), 199–250.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). **Does Job Corps Work? Impact Findings from the National Job Corps Study**. *The American Economic Review*, *98*(5), 1864–1886.

Torp, H., Raaum, O., Hernaes, E., & Goldstein, H. (1993). **The first Norwegian experiment**. In Karsten, J. & Madsen, P. K. (Eds.), *Measuring labour market measures: Evaluating the effects of active labour market policies. Copenhagen, Ministry of Labour*. Copenhagen: Ministry of Labour.

Viñas, V. (2009). **The European Union's Drive towards Public Policy Evaluation The Case of Spain.** *Evaluation*, *15*(4), 459–472.