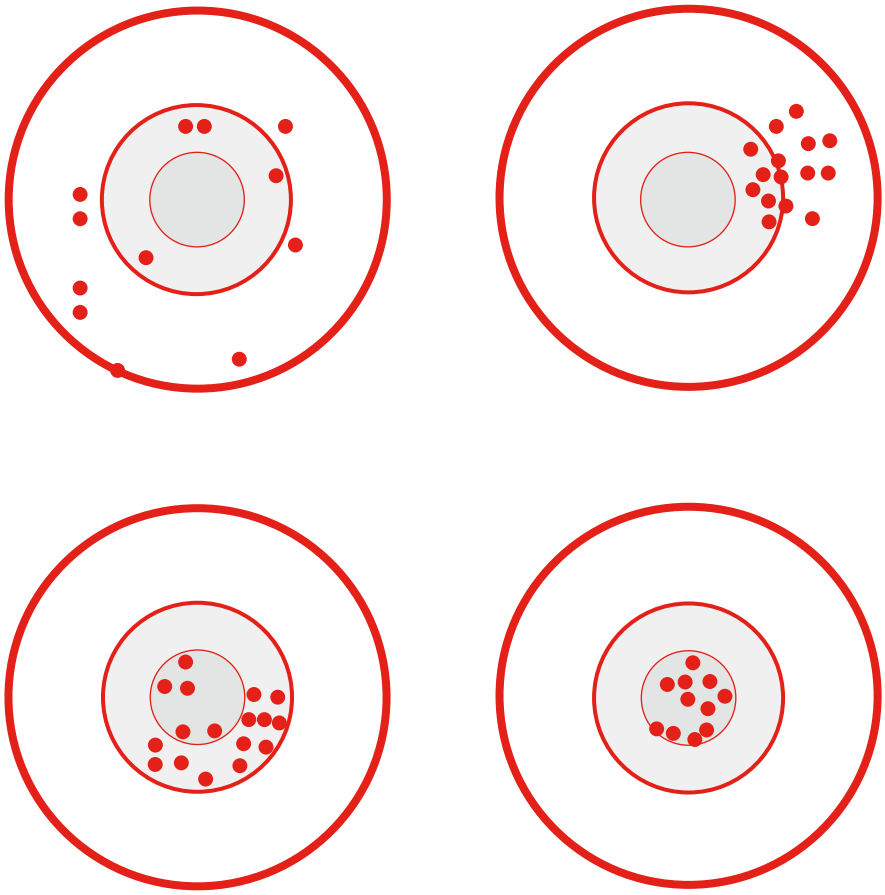
Guia pràctica 5
Avaluació d'impacte

Col·lecció Ivàlua de guies pràctiques sobre avaluació de polítiques públiques



ivàlua | Institut Català d'Avaluació de Polítiques Públiques

Institucions membres d'Ivàlua:

Generalitat de Catalunya Departament d'Economia i Finances

Diputació Barcelona

Universitat Pompeu Fabra

Fundació Jaume Bofill

Autors: Jaume Blasco, Analista d'Ivàlua
David Casado, Analista d'Ivàlua
Disseny: petitcomite.net
Impressió: Cevagraf, s.c.c.l.

Primera edició: Setembre de 2009

Dipòsit legal: B-39211-2009

ÍNDEX

1. INTRODUCCIÓ

Les administracions públiques s'esmercen contínuament a dissenyar i intentar millorar polítiques i programes, i dediquen cada any milers de milions d'euros a implementar-los. No obstant això, problemes com ara l'atur, el fracàs escolar, la sinistralitat a les carreteres o la degradació ambiental, tendeixen a persistir, la qual cosa planteja dubtes sobre l'efectivitat de les intervencions públiques que han de fer-los front. D'una part, aquest fet palesa que la tasca d'enfrontar-se als problemes socials és complicada, que en el millor dels casos mena a avenços lents, graduals i incomplets. De l'altra, que, encara que una intervenció pública sembli una gran idea i s'hi destinin molts recursos, el seu èxit no es pot donar mai per garantit *a priori*.

Sobre la base d'una anàlisi sistemàtica *ex-post*, **l'avaluació d'impacte** tracta, precisament, de determinar la capacitat que tenen les idees potencialment bones per solucionar els problemes socials. Un augment dels impostos sobre el tabac fa de debò que la gent fumi menys? Oferir desgravacions fiscals per als plans de pensions fa que la gent estalviï més per després de la jubilació? Incrementar les hores lectives a l'educació primària millora el rendiment escolar? Formar els aturats amb baixa qualificació augmenta la seva renda a mig termini? Atès que els problemes socials poden tenir conseqüències greus per a qui els pateix, i que els recursos per fer-los front són limitats, es tracta d'identificar i destriar les polítiques públiques que millor funcionen per donar-los solució o, com a mínim, per contenir-los.

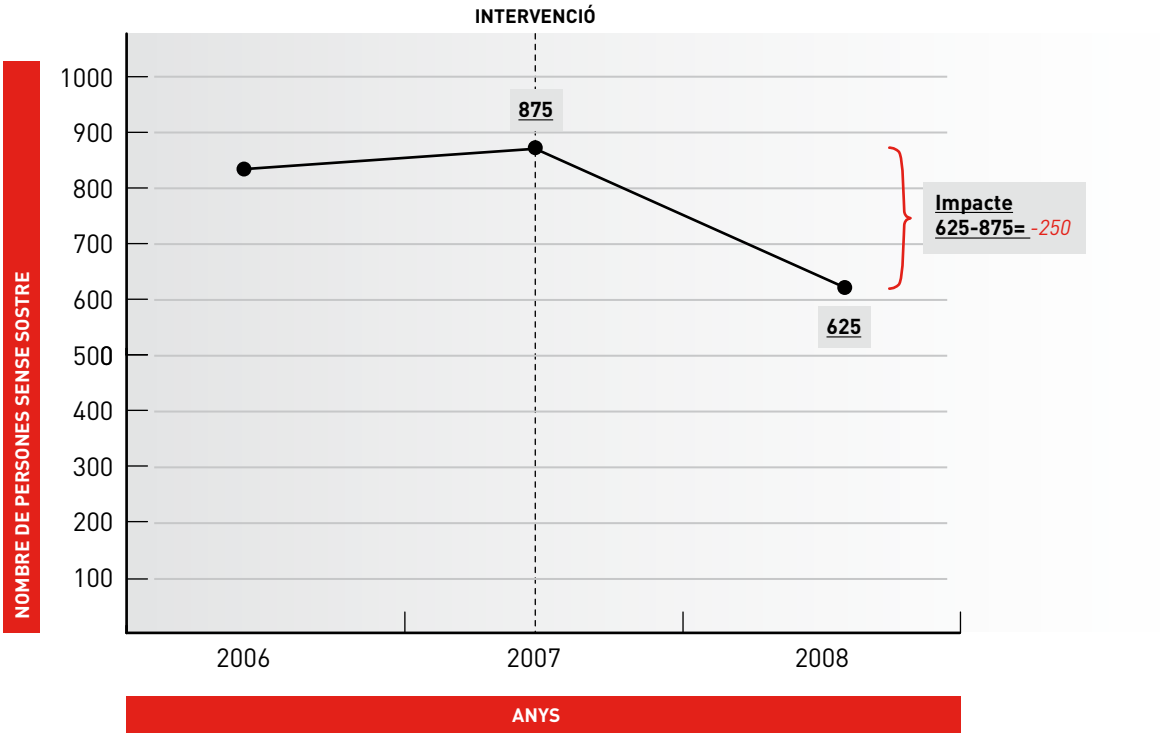
Tanmateix, com podem saber si les polítiques públiques realment funcionen? I si funcionen, quina és la magnitud del seu impacte? Massa sovint l'avaluació de les polítiques s'ha centrat exclusivament en els *inputs* o els *outputs*, és a dir, en els recursos que el programa utilitza o en allò que el programa fa. Però que 100 bombers hagin estat 24 hores abocant aigua sobre un foc ens diu poc sobre si han aconseguit apagar-lo. Tampoc el mer seguiment d'un problema ens diu gaire sobre l'impacte real de les polítiques públiques. Que un any s'hagin cremat la meitat d'hectàrees de bosc que l'any anterior no vol dir, necessàriament, que els bombers hagin fet millor la seva feina. Per tant, en una avaluació d'impacte no només volem saber si un problema millora o empitjora, sinó si la intervenció pública hi ha tingut alguna cosa a veure. Es tracta, en resum, d'establir si es pot atribuir o no (i en quina mesura) la *causa* del canvi en el problema a la intervenció pública. Fer-ho de forma convincent, com veurem, és una tasca laboriosa.

1.1. AVALUACIÓ D'IMPACTE: A LA RECERCA DE LA CAUSALITAT

Suposem que l'ajuntament d'una ciutat posa en marxa un programa nou d'atenció a les persones sense sostre que pernocten als carrers de la ciutat. La intervenció consisteix a oferir atenció personalitzada al carrer a les persones que no utilitzen els recursos municipals residencials, amb el propòsit que cada persona rebi sempre l'atenció del mateix treballador social. El programa és costós, perquè implica la contractació de nombrosos treballadors socials nous, però s'espera que ajudi a reduir considerablement el nombre de persones que es troben en la necessitat de dormir al ras. La teoria consisteix que, mitjançant aquest tipus d'atenció, el treballador social desenvoluparà una relació de confiança amb la persona sense sostre que li permetrà detectar millor quins problemes pateix, informar-la, orientar-la i acompanyar-la al recurs o servei més adequat en cada cas, i anar vencent les barreres i desconfiances que fan que les persones sense sostre es mostrin reticents a utilitzar aquests recursos. Se suposa que d'aquesta manera s'incrementarà el nombre de persones que entren en contacte amb el sistema d'atenció, la qual cosa constitueix un primer pas crític per poder proveir-les de l'assistència que necessiten i, en últim terme, permetre que aquestes persones accedeixin a solucions residencials estables en què puguin desenvolupar el seu projecte de vida amb el màxim d'autonomia personal possible. La idea sembla bona, però, funcionarà a la pràctica?

Imaginem, ara, que per mesurar l'impacte dels seus programes per a les persones sense sostre l'ajuntament realitza recomptes anuals de les persones que pernocten al carrer. Tal com mostra el gràfic 1, els recomptes realitzats amb anterioritat al nou programa indicaven que l'any 2007 hi havia 875 persones dormint als carrers de la ciutat, i que aquest nombre representava una petita variació respecte a l'any anterior. El recompte de l'any 2008, esperat amb expectació per poder estimar l'impacte del nou programa, revela que la població de persones sense sostre ha baixat fins a les 625 persones. En altres paraules, això implica una reducció de 250, gairebé un 30% comparat amb la població sense sostre de l'any anterior. A primera vista, sembla que l'impacte del programa hagi estat positiu i considerable. Tanmateix, podem considerar que aquesta conclusió és prou acurada?

Gràfic 1: Evolució de les pernoctacions al carrer



Per respondre aquesta pregunta hem de tenir en compte que, entre els anys 2007 i 2008, poden haver passat altres coses a part de la posada en marxa del nou programa. Per exemple, és possible que l'economia s'hagi expandit i ofereixi més oportunitats laborals fins i tot per a les persones de més baixa qualificació. També pot haver passat que els serveis de salut mental hagin endegat un nou programa en coordinació amb els serveis socials, que s'hagi mostrat especialment efectiu a prevenir que les persones amb malalties mentals greus i pocs recursos econòmics acabin al carrer. Igualment, és possible que el govern hagi endurit el control de l'entrada al país de nous immigrants, fent més difícil l'entrada al país per als immigrants indocumentats, que representen un sector de la població amb problemes més greus d'accés a l'habitatge. Aquests fenòmens, entre molts altres, podrien explicar, totalment o parcialment, el descens de la població sense sostre observada entre 2007 i 2008. La situació oposada és igualment factible: que en aquest mateix any les condicions econòmiques haguessin empitjorat, un programa d'atenció a les persones amb malalties mentals s'hagués suprimit i haguessin entrat a la ciutat molts més immigrants indocumentats que en anys anteriors. En aquest cas, la reducció de 250 persones respecte a l'any anterior estimada al gràfic 1 seria una clara subestimació de l'impacte real del programa.

La situació descrita en l'exemple és la més habitual en una avaluació d'impacte. Podem mesurar fàcilment un determinat fenomen, com ara la quantitat de persones que dormen al carrer, el nombre d'accidents a les carreteres, o la productivitat del sector de la fruita dolça per capturar l'impacte o *outcome* d'una intervenció pública que ens interessa avaluar. Però, malauradament per als avaluadors, succeeixen moltes altres coses més enllà de la mateixa

intervenció pública (com ara l'evolució de l'economia, els canvis en la meteorologia o la posada en marxa d'altres programes i polítiques) que tenen una influència notable sobre l'impacte que intentem observar i en compliquen l'avaluació. En conseqüència, avaluar l'impacte d'un programa implica ser capaç d'aïllar l'efecte del programa en relació amb tots aquests altres fenòmens que afecten el problema o situació que la intervenció pública pretén adreçar.

Aquesta constatació ens porta a introduir el que sembla un petit matís però que té, en realitat, una importància cabdal en l'avaluació d'impacte (i que, com veurem més endavant, és la principal font de maldecaps metodològics): la pregunta que l'avaluació d'impacte ha de respondre no és què ha passat després de posar en marxa una intervenció pública (moltes coses hi poden haver influït), sinó què ha passat *en comparació amb el que hauria passat si la intervenció no s'hagués dut a terme*. Lògicament, la diferència entre el que ha passat amb el programa i el que hauria passat sense el programa es pot atribuir únicament i solament al programa o, dit d'una altra manera, la diferència ha estat *causada* pel programa. I això és, precisament, el que cerca l'avaluació d'impacte: allò que el programa ha causat, i no allò que ha passat al mateix temps que el programa.

QUADRE 1
ASSOCIACIÓ NO VOL DIR CAUSALITAT

Una de les regles d'or presents a quasi tots els manuals d'estadística és no confondre associació amb causalitat. La diferència entre tots dos conceptes és senzilla. Suposem que, en un moment donat, observéssim en una població determinada que tenir els dits esgrogueïts i patir bronquitis crònica estan associats, és a dir, són característiques que tendeixen a presentar-se juntes en les mateixes persones. Vol dir això que la bronquitis crònica fa que la gent tingui els dits grocs? En realitat sabem que no és així, sinó que hi ha un tercer factor, que és fumar, que és una causa important tant que la gent tingui els dits grocs, com que pateixin bronquitis crònica. Per això tenir bronquitis i els dits grocs són fenòmens associats, però un no és la causa de l'altre. Tècnicament, es diu que l'associació que existeix entre tots dos fenòmens és *espúria*.

Tanmateix, desenredar causalitat i associació en el camp de les polítiques públiques no sempre és tan fàcil. Imaginem-nos que, entre la població escolar, estudiar en una escola concertada està associat a un millor rendiment acadèmic que fer-ho en una escola pública. Vol dir això que la titularitat de l'escola és la causa d'aquesta diferència i, per tant, que el concert escolar és una forma de provisió de l'educació més efectiva que la gestió pública directa? És possible, però no és segur. Una explicació alternativa és que els alumnes de l'escola concertada tendeixen a pertànyer a famílies d'un nivell socioeconòmic i formatiu superior que els de la pública, i que aquesta diferència en les característiques de l'alumnat és la causa real de la diferència en el rendiment escolar. De manera similar, que un ajuntament posi en marxa un programa d'atenció a les persones sense sostre i l'any següent baixi considerablement el nombre de persones que pernocten al carrer són fets associats, però no necessàriament un és la causa de l'altre. Com hem vist en l'explicació de l'exemple, hi ha molts altres motius plausibles, per la qual cosa és millor no extreure conclusions precipitades que ens puguin convertir en víctimes de la *fal·làcia causal*.

Quan observem una associació (per exemple, que participar en un programa està associat a una millora en un determinat *outcome*), és important tenir sempre present que la causalitat és una explicació possible, però no l'única. El repte de l'avaluació d'impacte és, justament, descartar explicacions alternatives per poder atribuir, de la forma més convincent possible, la causalitat del canvi observat a la intervenció pública.

1.2. EL CONTRAFACTUAL I L'ESTIMACIÓ DE L'IMPACTE D'UNA POLÍTICA PÚBLICA

Seguint l'argumentació del paràgraf anterior, l'impacte d'una intervenció pública es pot expressar en termes de la diferència entre dos nombres:

$$\text{IMPACTE} = Y_1 - Y_0$$

On:

- Y_1 són els *outcomes* que s'han esdevingut amb la intervenció pública.
- Y_0 són els *outcomes* que s'haurien esdevingut en absència de la intervenció pública, que de forma tècnica (i més breu), s'anomenen el *contrafactual*.

De forma general, Y_1 és un nombre relativament fàcil d'estimar. Normalment, fent ús de registres administratius, mitjançant una enquesta, realitzant un recompte (com en l'exemple) o amb qualsevol altra tècnica d'observació, podem estimar què ha passat amb els *outcomes* d'interès un cop s'ha implementat el programa. Per exemple, podem arribar a saber, sense gaires dificultats, quants aturats han trobat feina després de participar en un curs de formació, quantes patents s'han registrat en el marc d'un programa de subvencions R+D+I, o com han evolucionat les rendes dels agricultors després d'un programa de suport a la tecnificació d'un determinat tipus de conreu.

Per contra, estimar Y_0 són figues d'un altre paner. De fet, construir un contrafactual apropiat és, de bon tros, la tasca més complicada de l'avaluació d'impacte. El motiu d'aquesta dificultat és, senzillament, que el món no pot estar en dos estats al mateix temps: una ciutat no pot haver implementat un programa i no implementar-lo al mateix temps, igual que una empresa no pot haver rebut una subvenció R+D+I i no rebre-la alhora. Si el programa s'ha implementat, mai no podrem arribar a observar què hauria passat si no s'hagués dut a terme. Per tant, mentre que l'estimació d'Y_1 respon a una mesura basada en l'observació de la realitat, l'estimació d'Y_0 és sempre una declaració hipotètica sobre com creiem que hauria estat el món en absència del programa.

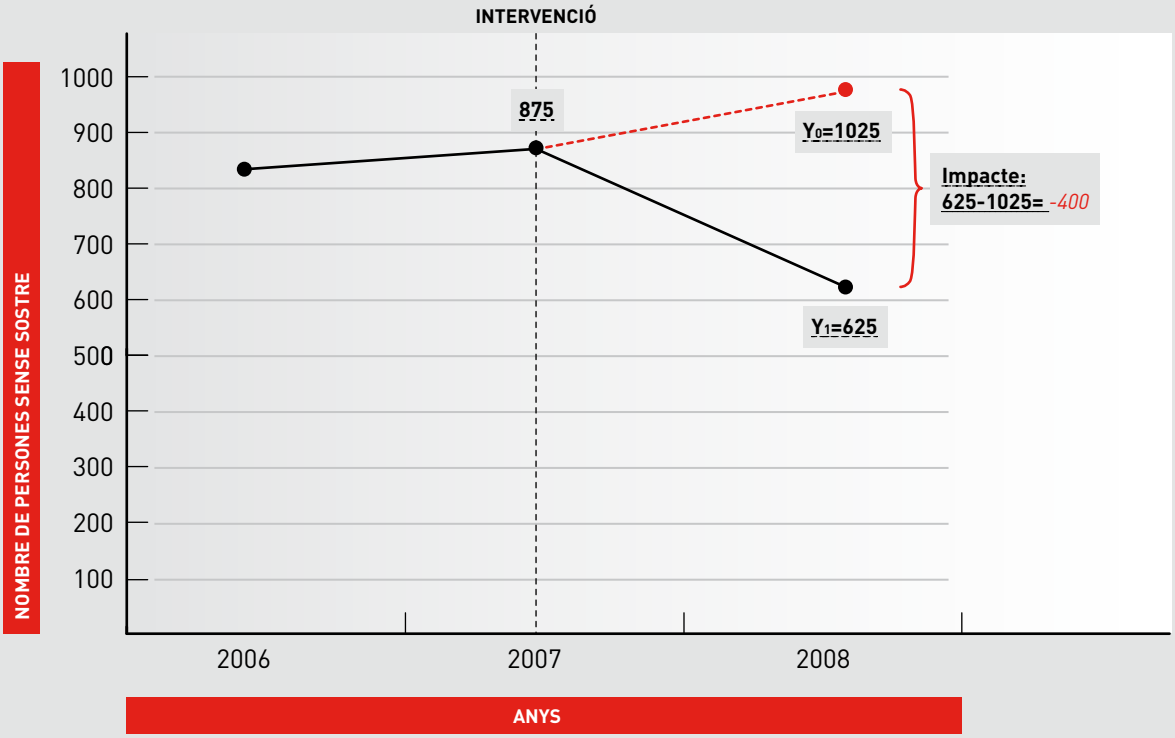
La qüestió llavors esdevé: com ens ho fem per formular una hipòtesi contrafactual? El concepte en si mateix no ens hauria d'espantar car la vida quotidiana és plena d'exemples d'aquest tipus d'hipòtesis: "si hagués estudiat més, hauria aprovat les oposicions"; o bé "si no m'hagués hipotecat, ara no aniria tan escanyat". El repte de l'avaluació de polítiques, però, és arribar a construir una hipòtesi que no només sembli realista sinó que, a més, permeti quantificar amb precisió què hauria passat en absència del programa, ja que necessitem un nombre Y_0 amb el qual poder realitzar la resta (Y_1 - Y_0) que ens porta a estimar l'impacte del programa.

Per fer-ho, l'estratègia sol consistir a substituir el contrafactual, que per definició és no observable, per un escenari de comparació observable. Per exemple, suposem que el Departament d'Educació endega un programa que consisteix a atorgar autonomia de gestió a les direccions de determinats centres escolars, amb l'objectiu de millorar la qualitat de l'educació i, en últim terme, el rendiment dels alumnes. Mesurar Y1 és fàcil: es tracta de mesurar quines qualificacions han tret els nens i nenes d'aquests centres escolars, diguem que un any després del canvi en el model de gestió. Quina pot ser la hipòtesi contrafactual? Suposem que a la xarxa de centres escolars hi ha escoles *de característiques similars a les que han participat en el programa*, que romanen sota el règim de gestió ordinari. Podem mesurar les qualificacions dels alumnes d'aquests centres similars i formular la següent hipòtesi contrafactual: si les escoles que han participat en el programa no ho haurien fet (contrafactual no observable), les qualificacions que haguessin tret els seus alumnes serien les mateixes que han tret els alumnes de les escoles de característiques similars que no hi han participat (escenari de comparació observable).

QUADRE 2
LA MESURA DE L'IMPACTE AMB UNA HIPÒTESI CONTRAFACTUAL

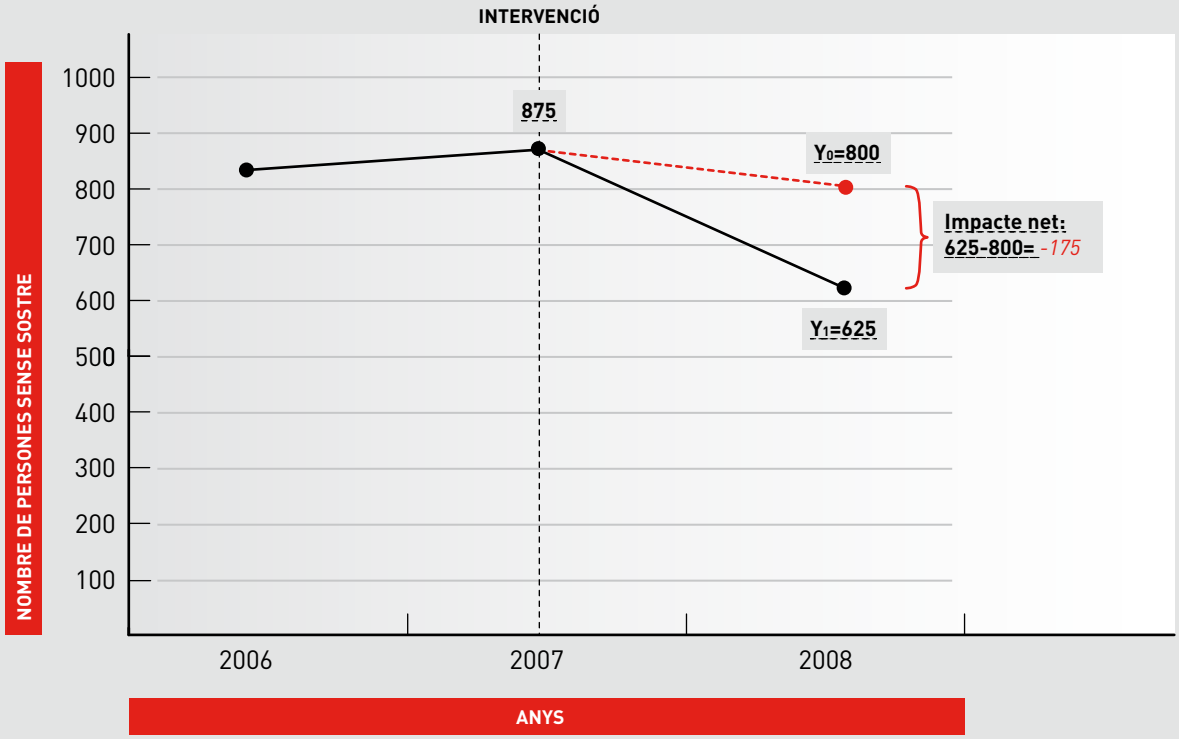
Els gràfics 2 i 3, tornant a l'exemple del programa d'atenció a les persones sense sostre, representen amb una línia vermella dos possibles contrafactuals. El primer es basa en l'estimació que, en absència de la intervenció, el nombre de persones sense sostre hauria augmentat (això correspondria, per exemple, a un escenari de més atur, pitjors serveis a les persones amb malaltia mental i més immigració indocumentada). D'altra banda, en l'estimació del contrafactual del gràfic 3 s'assumeix que el nombre de persones hauria baixat igualment en absència del programa (a causa, per exemple, d'un escenari de menys atur, millors serveis i menys immigració). Noteu que Y_1 no varia en els dos gràfics: el programa es va endegar i el nombre de persones que pernoctaven al carrer després de la implementació es va observar i mesurar. Per tant, la divergència de la magnitud de l'impacte en un i altre gràfic (400 persones en el gràfic 2 i 175 en el 3) es deu exclusivament al fet que l'estimació d'Y_0 (el contrafactual) és diferent.

Gràfic 2: Evolució de les pernoctacions al carrer amb estimació del contrafactual (I)



Quadre 2 (cont.)

Gràfic 3: Evolució de les pernoctacions al carrer amb estimació del contrafactual (II)



La bibliografia estadística i economètrica és plena d'estratègies per identificar el contrafactual de programes i polítiques públiques, i en el capítol 3 d'aquesta guia exposarem les d'ús més freqüent. Comprovarem que el principal repte d'aquestes **estratègies d'identificació** rau a trobar unitats (escoles, persones, barris, etc.) que compleixin la condició de tenir *característiques similars* a les que han participat en el programa. Això és perquè, generalment, si una persona participa en un programa i una altra no, i si un barri rep una subvenció i un altre no, és perquè són diferents en alguna característica rellevant. Les estratègies d'identificació del contrafactual fan tots els possibles per controlar aquestes diferències, amb l'agreujant que mentre que algunes són observables, altres no ho són. Per exemple, podem trobar aturats que s'assemblin als que han participat en un curs de formació quant al nivell formatiu previ, la història laboral, l'edat i altres característiques similars recollides en una base de dades, però no per altres factors rellevants, com ara la motivació per trobar feina, l'estat anímic, etc.

Les metodologies per a l'avaluació d'impactes que presentarem al capítol 3 s'adeqüen a diferents tipus de programes i circumstàncies de l'avaluació, i no n'hi ha cap d'universalment superior. La selecció de l'estratègia més adequada requerirà, en cada cas, una anàlisi prèvia de les característiques de la intervenció pública que la justifiqui, especialment sobre els objectius del programa, el procediment de selecció dels participants, el procés d'implementació i les fonts de dades disponibles. Abans d'exposar els diferents mètodes per a l'avaluació d'impacte, el capítol 2 fa referència als passos preliminars per enfocar el disseny de l'avaluació i que guiaran l'elecció del mètode més adequat.

QUADRE 3
LA ROBUSTESA DE LES HIPÒTESIS CONTRAFACTUALS

Les estratègies d'identificació del contrafactual són hipòtesis sobre situacions que mai no s'esdevindran ja que, com hem dit, és impossible que una persona que ha participat en un programa no hi hagi participat al mateix temps. Per tant, totes les estratègies tenen en comú que no poden ser testades empíricament, és a dir, mai no podrem comprovar *a posteriori* si eren correctes o falses. Tot el que podem fer és valorar si la hipòtesi contrafactual sembla més o menys realista i argumentar sobre els motius pels quals creiem que es tracta (o no) d'una hipòtesi plausible. De fet, les controvèrsies sobre les avaluacions giren gairebé sempre sobre la robustesa de la hipòtesi contrafactual, és a dir, sobre com d'adequat és l'escenari de comparació identificat.

La bibliografia és plena d'exemples de programes o polítiques en què diferents estratègies d'identificació en l'avaluació han menat a estimacions de l'impacte molt diferents. Per exemple, les avaluacions sobre la cooperació financera internacional amb els països en desenvolupament han tendit a no detectar cap impacte significatiu sobre el creixement econòmic dels països receptors. Tanmateix, l'any 2000, els economistes del Banc Mundial Criag Burnside i David Dollar van publicar un article en què introduïen una novetat en aquest tipus d'avaluacions: l'efectivitat dels ajuts financers podria dependre de la qualitat de les institucions i polítiques fiscals, monetàries i comercials del país receptor. Efectivament, la seva avaluació indicava que si la comparació es feia només entre països amb *bona governança*, l'impacte de l'ajut financer era positiu i estadísticament significatiu. En canvi, entre països amb institucions i polítiques deficients, l'impacte de la cooperació financera era nul. Aquesta avaluació va ser altament influent, ja que va menar diverses institucions a condicionar la seva cooperació financera a l'adopció, per part dels països receptors, de les polítiques i les institucions identificades com a adequades en l'article de Burnside i Dollar.

Posteriors articles i avaluacions han posat en solfa l'estratègia d'identificació del contrafactual emprat en l'esmentat article, amb la qual cosa la pregunta d'avaluació de fons —*els països que reben ajut financer internacional, es desenvolupen econòmicament més de pressa que si no en rebessin?* — encara sense una resposta clara.

2. PASSOS PRELIMINARS PER DISSENYAR UNA AVALUACIÓ D'IMPACTE

2.1. ÉS OPORTÚ AVALUAR ELS IMPACTES DEL PROGRAMA?

L'avaluació d'impacte és, en certa manera, la reina de les avaluacions. Malgrat la innegable importància de realitzar una avaluació de necessitats per caracteritzar adequadament el problema que es vol adreçar, de ponderar bé el disseny de la intervenció i assegurar-se que és robust i coherent amb el coneixement que les ciències socials atresoren, i d'avaluar el procés d'implementació per detectar dificultats imprevistes i desviacions respecte de les previsions, pocs moments són tan emocionants tant per als gestors dels programes com per als avaluadors com el d'intentar respondre la pregunta *"funciona?"*. Pot existir, en conseqüència, la temptació de fer-se aquesta pregunta prematurament, quan les condicions no són encara adequades per poder realitzar una avaluació d'impacte, o bé quan seria més aconsellable i rellevant efectuar un altre tipus d'avaluació. Els requisits per dur a terme una avaluació d'impacte són aquests:

1. **El programa ha de ser estable.** Per poder avaluar els impactes d'una intervenció pública és molt convenient que aquesta intervenció hagi romàs sense gaire variacions durant un cert temps, ja que, altrament, es fa difícil determinar sobre quina de les múltiples versions del programa s'han d'estimar els impactes. A més, en programes inestables o *volàtils,* és molt possible que els resultats de l'avaluació d'impacte siguin irrellevants des del mateix moment en què es coneguin perquè la versió avaluada no coincideix amb la que s'està implementant en aquell moment. L'estabilitat del programa sol ser més baixa quan el programa és relativament nou, ja que en les primeres fases de la implementació és habitual que es produeixi un cert procés d'ajustament del programa per assaig i error. En aquestes circumstàncies, una avaluació d'implementació que permeti analitzar de forma sistemàtica què està passant i detectar quines correccions escauen sol ser més útil que una avaluació d'impacte. No obstant això, l'excepció a aquesta regla la constitueixen els programes pilot, que si bé per definició són sempre nous, es mantenen estables i fidels al seu disseny original, precisament perquè el seu objectiu és avaluar-ne l'efectivitat.

2. **Cal haver descrit una teoria del canvi coherent.** Com qualsevol altre tipus d'avaluació, l'avaluació d'impacte requereix que prèviament s'hagin identificat els objectius genuïns de la intervenció (altrament no és possible determinar quins són els impactes que s'han estimar) i una teoria del canvi que uneixi, de forma mínimament plausible, les activitats i els productes del programa amb els impactes que es pretenen assolir (ja que, si esperar impactes positius es demostra poc realista, és preferible treballar per millorar el disseny de la intervenció que per avaluar-ne els improbables impactes). Dit en altres paraules, abans de l'avaluació d'impacte (o en el marc de l'avaluació d'impacte) és necessària una mínima avaluació del disseny.

3. **Cal tenir un coneixement adequat del procés d'implementació.** L'interès per saber si un programa funciona o no sol anar acompanyat de l'interès per saber per què funciona, per la qual cosa les avaluacions d'impacte sovint es fan juntament amb avaluacions de la implementació. Però fins i tot si el nostre interès se centra estrictament a mesurar els impactes de la intervenció, un mínim coneixement del procés d'implementació és necessari per interpretar els resultats d'una avaluació d'impacte i transformar-los en recomanacions de millora. Així, si una avaluació d'impacte conclou que un programa no té cap impacte significatiu, és possible concloure que la teoria de l'impacte que uneix els *outputs* amb els *outcomes* era equivocada (vegeu: Ivàlua. Guia pràctica, 3[1]), o que el programa mai no es va arribar a implementar com estava previst i els *outputs* previstos mai no es van arribar a generar, ja sigui per desviacions respecte a disseny o perquè la teoria del procés era impossible de portar a la pràctica. Fins i tot si els resultats de l'avaluació d'impacte són positius, comprovar que el procés d'implementació s'ha produït d'acord amb les previsions reforça la conclusió que el programa és la causa dels impactes.

4. **Els impactes s'han d'haver pogut produir.** Són rares les intervencions públiques que produeixen impactes immediats, per la qual cosa és necessari que transcorri un cert temps des de la implementació de la intervenció abans de poder-ne detectar l'impacte. A les pàgines que segueixen veurem que una de les decisions a prendre a l'hora de dissenyar una avaluació d'impacte és triar el moment més adequat per fer la mesura de l'impacte, ja que és possible que alguns efectes triguin a esdevenir-se, tendeixin a acumular-se o es dissipin amb el temps. Si, donat el tipus d'intervenció, sabem de bon principi que aquest moment no ha arribat encara, serà preferible posposar l'avaluació i esperar que els impactes hagin pogut tenir lloc.

2.2. A QUÈ ENS REFERIM QUAN PARLEM D'*OUTCOMES*?

Al llarg d'aquesta sèrie de guies metodològiques hem repetit en diverses ocasions que les polítiques públiques tenen la seva raó de ser en l'existència d'un problema o situació social insatisfactòria, i que els objectius de la política pública han de fer referència al canvi que la intervenció pública pretén induir sobre aquest problema o situació. Sembla, doncs, que la definició dels *outcomes* amb els quals mesurarem l'impacte s'hauria de derivar de forma bastant directa dels objectius del programa, ja siguin els declarats formalment o els identificats en l'elaboració de la teoria del canvi de la intervenció. Per exemple, si l'objectiu d'un programa és la reducció de la sinistralitat a les carreteres, sembla que la definició dels *outcomes* hauria de capturar de la millor manera possible el fenomen de la *sinistralitat a les carreteres*. Tanmateix, la tasca d'identificar els *outcomes* i la forma de mesurar-los rarament és directa i sol requerir la presa d'algunes decisions sobre què, com i quan mesurar.

En primer lloc, cal tenir present que algunes intervencions públiques tenen objectius múltiples. Per exemple, la reducció de la velocitat màxima en els accessos a Barcelona té per objectiu reduir la contaminació i reduir els accidents; i el Programa Interdepartamental de Rendes Mínimes d'Inserció té per objectiu, com el seu nom indica, elevar la renda i inserir en el mercat laboral les persones beneficiàries de la prestació. Si aquest és el cas del programa que hem d'avaluar, és necessari seleccionar l'objectiu sobre el qual ens interessa avaluar els impactes, o si decidim avaluar-ne més d'un, prendre consciència en la planificació de l'avaluació que els recursos necessaris (temps, finançament, etc.) es multiplicaran.

D'altra banda, alguns objectius són multidimensionals. Fins i tot si la intervenció té un únic objectiu, o si n'hem triat un de sol sobre el qual volem realitzar l'avaluació d'impacte, les maneres com podem arribar a definir aquest impacte solen ser múltiples. Suposem, per exemple, que volem capturar el fenomen de la sinistralitat a les carreteres: podem mesurar el nombre d'accidents, el d'accidents amb ferits o morts, o bé directament el nombre de ferits o de morts en accident de trànsit. Per contra, si volem capturar el fenomen de la inserció laboral, que sol ser l'objectiu de les polítiques actives d'ocupació, les opcions es multipliquen: ens pot interessar si la persona ha trobat una feina dins d'un període de temps, o bé intentar capturar la retenció de la feina, és a dir, mesurar si la persona manté la feina al cap d'un temps determinat, o bé quants dies en total ha treballat al llarg d'aquest període de temps. Igualment, és possible que el nostre interès en la inserció laboral sigui instrumental, amb la qual cosa la dimensió que realment ens resulta rellevant és la variació en la renda o bé l'increment en el benestar subjectiu derivats de la inserció laboral. En els termes que empràvem en la Guia 3 sobre avaluació de disseny, la consecució d'alguns objectius implica l'assoliment d'una seqüència prèvia d'impactes (per exemple, trobar feina, retenir-la, fet que incrementa la renda i, en últim terme, el benestar), que anomenàvem *estructura d'impactes*. Abans d'iniciar l'avaluació és precís decidir quina (o quines) de les múltiples dimensions que constitueixen aquesta estructura és la més rellevant per al propòsit de la nostra avaluació.

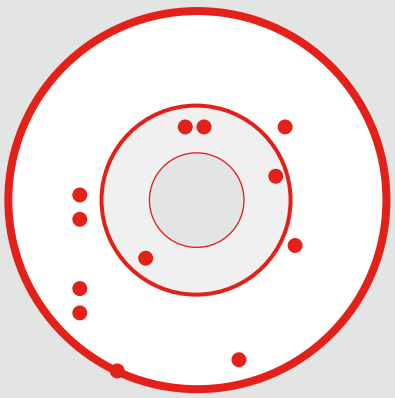
QUADRE 4
MESURES!

"Suposem que us contracto per mesurar un elefant. Pot semblar una tasca prou clara, però penseu-hi un minut. Heu de mesurar-ne pes? L'alçada? La llargada? El volum? La intensitat del seu color gris? El nombre i profunditat de les seves arrugues? O potser la proporció del dia que dorm? Per poder mesurar aquesta criatura necessiteu seleccionar una o unes quantes característiques entre diverses possibilitats. L'elecció dependrà del vostre propòsit a l'hora de mesurar, o més aviat del meu, ja que us he contractat. Si jo fos el responsable del transport ferroviari de mercaderies necessitaria conèixer l'alçada, la longitud i el pes de l'elefant. Però si fos un taxidermista, estaria més interessat en el seu volum i arrugues. Com a domador, em preocuparia més la proporció del dia que dorm. Com a productor de pells d'animal sintètiques, voldria saber el to exacte del gris. Vosaltres, veient l'oportunitat de mantenir-vos en nòmina, insistiríeu segurament en el fet que no puc entendre el meu elefant si no en conec la variació estacional de la temperatura corporal".

STONE, D. *Policy Paradox, The Art of Political Decision Making, 2002* [Traducció pròpia]

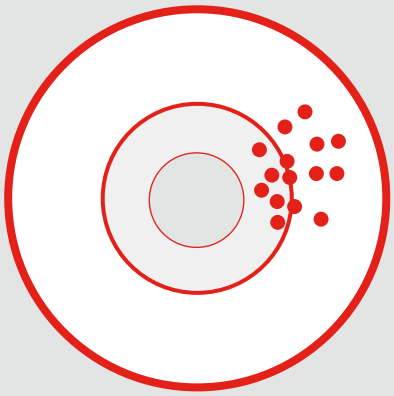
Per contra, alguns impactes són especialment difícils de mesurar perquè els objectius fan referència a constructes particularment intangibles com, per exemple, incrementar l'autonomia personal dels participants d'un programa d'atenció a les persones sense sostre. En aquest cas, la dificultat no és tant seleccionar una dimensió entre les diverses que constitueixen un objectiu, sinó arribar a mesurar un fenomen que, per la seva naturalesa, sembla immesurable. En aquestes situacions, la decisió rau entre triar una mesura preexistent (hi ha, en aquest sentit, una literatura especialitzada en el desenvolupament de mesures per als fenòmens socials més diversos, des del desenvolupament cognitiu a l'estrès laboral, passant per la felicitat i la percepció de seguretat a la via pública) o bé crear-ne una de nova ajustada a les especificitats de la nostra avaluació. En general, sol ser preferible triar una mesura preexistent, ja que això implica que algú n'ha comprovat amb anterioritat la **fiabilitat** (és a dir, que si la mesura s'empra en diverses ocasions els resultats són coherents), i perquè l'ús d'una mesura estandarditzada facilita la posterior comparació de resultats amb altres avaluacions. A més, l'esforç de localitzar una mesura **vàlida** per a la nostra avaluació a la bibliografia (o sigui, que capturi satisfactòriament el nostre fenomen d'interès) sol ser substancialment menor que el de desenvolupar i testejar qüestionaris per elaborar-ne una de pròpia.

QUADRE 5
ELS CONCEPTES DE VALIDESA I FIABILITAT DE LA MESURA DE L'IMPACTE



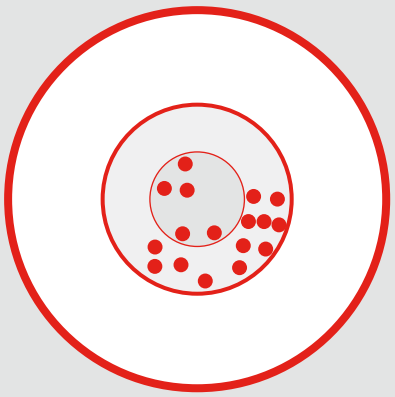
NI VÀLIDA NI FIABLE

La mesura no captura el fenomen d'interès, i els diferents intents de mesurar el fenomen donen resultats dispars.



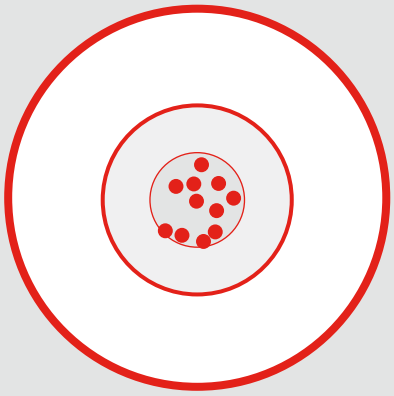
FIABLE PERÒ NO VÀLIDA

La mesura no captura el fenomen d'interès, però intents repetits de mesurar el fenomen donen sempre resultats molt similars (però equivocats)



RELATIVAMENT VÀLIDA PERÒ POC FIABLE

La mesura captura relativament bé el fenomen d'interès, però els diferents intents de mesurar el fenomen donen resultats massa dispars.



FIABLE I VÀLIDA

La mesura captura el fenomen d'interès, i els intents repetits de mesurar el fenomen donen sempre resultats molt similars.

Font: Adaptació de VARKEVISSER, C. M.; PATHMANATHAN, I.; BROWNLEE, A. Designing and conducting health systems research projects. World Health Organization / International Development Research Centre, 2003.

Per acabar, hem de tenir en compte que definir els impactes no només implica especificar què mesurem i com ho mesurem, sinó també quan ho mesurem. Aquesta qüestió té una especial importància, ja que diferents moments de mesura poden menar a conclusions diferents sobre els impactes del programa, ja que mentre que alguns impactes impliquen processos lents i poden trigar a esdevenir-se, d'altres poden produir-se ràpidament però no mantenir-se en el temps. En aquest sentit, es tracta de determinar quin és el moment més rellevant per fer-se la pregunta *què ha passat en comparació al que hauria passat si la intervenció no s'hagués dut a terme?* Així, en un programa de suport als funcionaris perquè deixin de fumar, l'impacte pot ser fulgurant una setmana després de començar, però sembla més rellevant conèixer l'impacte un any després, ja que és probable que una part dels qui ho deixen inicialment acabin per recaure-hi. Per contra, una intervenció per protegir l'hàbitat d'una espècie amenaçada pot no tenir impactes apreciables en un principi, però molt notables al cap de tres anys, un cop la població ha tingut prou temps per augmentar sota les noves condicions. En qualsevol cas, el temps de mesura s'ha de definir amb precisió: no es pot parlar de curt o llarg termini, sinó que cal decidir, amb exactitud, si ens referim a sis, dotze, divuit o vint-i-quatre mesos després del programa.

2.3. QUÈ VOL DIR PARTICIPAR EN EL PROGRAMA?

Alguns conceptes i mètodes de l'avaluació quantitativa d'impactes estan parcialment importats de les ciències mèdiques. Per exemple, en l'assaig d'un medicament s'administra una píndola a algunes persones que reben el nom de grup de tractament o casos i un placebo a les persones que constitueixen el grup de control, o simplement *controls*. L'efecte del medicament s'infereix de la diferència en l'evolució de la patologia o el símptoma de torn entre un grup i un altre. De forma anàloga, els mètodes per a l'avaluació d'impacte solen comparar un grup de *tractament* (les escoles, persones, barris, etc. que han participat en un programa) amb un grup de comparació o control (integrat pels qui no hi han participat) que serveix per controlar el contrafactual.

No obstant això, participar en un programa sol ser un concepte bastant més imprecís que empassar-se una píndola. Mentre que amb un medicament no hi ha situacions intermèdies (sabem si l'hem administrat i en quina dosi), la participació en un programa pot voler dir coses molt diferents. Per exemple, que un barri participi en la *Llei de Barris* vol dir que ha rebut finançament per fer determinades actuacions que són gestionades d'una determinada manera. La quantitat de finançament, el tipus d'actuacions i la forma com s'han gestionat canvia d'un barri a un altre, per la qual cosa que un barri hagi *participat* en la *Llei de Barris* es correspon amb situacions molt diverses. Igualment, que una persona hagi participat en un curs de formació ocupacional pot voler dir des de que ha assistit al 100% de les classes d'un curs de jardineria de 80 hores, fins que ha assistit al 50% de les classes d'un curs de mediació comunitària de 20 hores. En síntesi, el *tractament*, en el cas de les polítiques públiques,

pot ser molt heterogeni i planteja una qüestió a resoldre: *de què*, exactament, volem estimar l'impacte?

L'**heterogeneïtat del tractament** varia segons el tipus d'intervenció pública. Si el nivell de variabilitat en el que significa haver participat o haver-se beneficiat de la intervenció pública és important, podem prendre alguna de les següents mesures per adreçar-la:

- Imposar restriccions sobre la definició de *participació* (per exemple, només considerarem que una persona ha participat en un curs de formació ocupacional si ha assistit a un mínim del 80% de les classes d'un curs d'un mínim de 30 hores).

- Desagregar l'avaluació segons el tipus de participació (per exemple, es pot avaluar per separat l'impacte de la formació en jardineria del de mediació comunitària).

- Assumir l'heterogeneïtat del tractament com una característica de la intervenció pública, tenint sempre present que s'està inferint un impacte promig de participacions que en realitat són diverses.

2.4. PER A QUI VOLEM DETECTAR ELS IMPACTES?

Molt sovint les intervencions públiques s'adrecen a una població diana força heterogènia. Per exemple, la *Llei de Barris* beneficia des de barris de grans zones urbanes fins d'altres de ciutats petites, alguns cada cop més deshabitats i d'altres excessivament poblats. De forma similar, els programes d'atenció a les persones sense sostre atenen des de persones amb malalties mentals a immigrants acabats d'arribar amb cap altre problema que la manca de recursos i una xarxa social de suport, des de persones analfabetes a llicenciats universitaris. Donada aquesta diversitat en la població diana, no és d'estranyar que els programes puguin ser efectius per a determinats tipus de beneficiari mentre que no ho són per a d'altres. En aquest context d'**heterogeneïtat dels impactes**, estimar **impactes** promitjos per a tots els beneficiaris pot fer concloure que un programa és relativament inefectiu per a la majoria de persones quan en realitat és molt efectiu per a un subgrup. En aquest cas, no es tractaria tant de descartar el programa com de mantenir-lo només per a aquells per a qui és efectiu i reformar-lo per als altres.

Si les dades disponibles ho fan possible, l'avaluació d'impacte permet no només saber si el programa *funciona*, sinó *per a qui funciona* mitjançant la desagregació de les estimacions d'impacte del programa per a diferents subgrups de població. Això permetria escatir, per exemple, si el carnet de conduir per punts és més efectiu per reduir la sinistralitat a les carreteres per als conductors joves o els de mitjana edat, per als infractors reincidents o per als ocasionals, o per als desplaçaments d'oci o els de treball.

En preparar el disseny d'una avaluació d'impacte, és important identificar quins són els subgrups de població (per gènere, grups d'edat, tipus de problemàtica inicial, etc.) per als quals és rellevant realitzar una anàlisi desagregada.

QUADRE 6
LES DECISIONS METODOLÒGIQUES EN EL PROCÉS DE DISSENY DE L'AVALUACIÓ

Dissenyar una avaluació d'impacte implica prendre decisions constantment: la definició de l'impacte, el moment de mesura, la concreció del que significa participar en el programa, la desagregació de l'anàlisi per subgrups o l'elecció del mètode per identificar el contrafactual no són passos automàtics sinó que impliquen triar una alternativa entre vàries.

Cadascuna d'aquestes decisions implica haver de resoldre una disjuntiva. D'una part, augmentar la complexitat de l'anàlisi (escollir més d'una definició d'impacte i moment de mesura, desagregar l'anàlisi en diversos graus de participació i subgrups de beneficiaris, o avaluar el programa amb més d'una metodologia) permet obtenir informació més detallada i assolir conclusions més robustes. De l'altra, incrementa el temps i recursos necessaris per dur a terme l'avaluació (de vegades fins a fer-la inabastable) i complica la comunicació dels resultats. En conseqüència, fins i tot si decidim que un cert grau de complexitat és assumible, és inevitable haver de renunciar a algunes mesures de l'impacte, nivells de desagregació i aproximacions metodològiques.

Tot i que, idealment, aquestes renúncies es fan sobre criteris de menor rellevància, la presa de decisions implica de vegades un cert grau d'arbitrarietat. Pot resultar difícil justificar per què mesurem la situació laboral al cap de 12 mesos i no de 24, desagreguem l'anàlisi per vegueries i no per grups d'edat, o per què hem triat un mètode determinat enlloc d'un altre, fins a generar una certa sensació que la foto que estem oferint sobre el rendiment del programa és incompleta.

Malgrat els dubtes que es plantegin en la presa d'aquestes decisions metodològiques, el més important és prendre-les amb diligència perquè l'avaluació pugui estar acabada a temps per ser rellevant, i fer constar sempre sota quina definició d'impacte i sota quines hipòtesis de partida hem arribat a la conclusió que el programa és efectiu o no.

2.5. A QUÈ ENS REFERIM, EXACTAMENT, QUAN PARLEM DE CONTRAFACTUAL?

Tal com hem explicat anteriorment, l'avaluació d'impacte requereix identificar un escenari contrafactual amb el qual estimar els *outcomes* que s'haurien produït en absència de la intervenció pública. Tanmateix, el concepte de contrafactual és excessivament ambigu i requereix ser precisat abans d'avançar en el disseny metodològic de l'avaluació:

- D'una part, les intervencions públiques rarament constitueixen el primer intent d'adreçar un problema, sinó que es tracta d'una reforma respecte d'un programa anterior. En aquest context, el contrafactual *és el que hauria passat si haguéssim continuat amb el programa antic.*

- De vegades, però, el programa pot ser genuïnament nou, o bé pot ser del nostre interès estimar l'impacte en relació amb l'absència de qualsevol intervenció pública. En aquestes situacions, el contrafactual esdevé *el que hagués passat si no hi hauria cap programa en funcionament.*

- Per acabar, de vegades, per a un mateix objectiu, hi ha diversos programes en funcionament, o bé n'hi ha un que funciona amb diferents variants o models d'implementació (per exemple, amb provisió pública directa en uns llocs i externalitzada en uns altres), amb la qual cosa l'interès de l'avaluació és valorar l'efectivitat d'un programa o model respecte dels altres. En aquestes situacions, el contrafactual es pot definir en qualsevol de les dues versions anteriors, depenent de la pregunta d'avaluació i l'aproximació metodològica per donar-li resposta.

2.6. DE QUINES DADES DISPOSEM PER FER L'AVALUACIÓ D'IMPACTE?

La disponibilitat de dades determina, a la pràctica, moltes de les decisions sobre el disseny d'una avaluació d'impacte. No sempre podem definir els *outcomes* com voldríem sinó que ens veiem forçats a definir-los de la millor manera que podem amb les dades de les quals disposem. El mateix passa amb la definició del que significa participar en el programa, la identificació dels subgrups d'interès o la selecció de l'estratègia metodològica per controlar el contrafactual. La disponibilitat de dades és el major determinant de la feina de l'avaluador, com el solar i l'entorn ho són per a l'arquitecte.

Aquestes limitacions es deuen al fet que, en general, és preferible treballar amb dades preexistents provinents de registres administratius. En efecte, si definim els *outcomes*, els subgrups i el tractament de manera que els puguem extreure de dades preexistents, l'avaluació és molt més ràpida i barata que si hem de dur a terme una enquesta per generar dades noves. A més, si usem dades administratives, la mostra amb què treballarem serà molt més gran que si fem una enquesta, amb la qual cosa les estimacions seran molt més precises. D'altra banda, ens estalviarem els biaixos de no resposta que pateixen les enquestes i que compliquen el tractament estadístic de les dades (Purdon, 2002). No obstant això, realitzar una enquesta no sempre és una opció a descartar. De vegades per estimar l'impacte d'un programa necessitem saber què ha passat amb els participants un temps després que hagin abandonat el programa, quan ja no se'n fa un seguiment en els registres. La disjuntiva és sempre entre el cost, el temps i les limitacions que implica generar dades noves mitjançant una enquesta, i l'avantatge de poder recollir tota la informació que ens interessa, i de la manera en què més ens interessa.

Tanmateix, les limitacions que la qualitat i el contingut dels registres administratius imposen sobre les avaluacions no han de ser considerades com un designi immutable. D'acord amb un reconegut economista i avaluador del Banc Mundial, que les avaluacions d'impacte siguin *ex-post* per definició no vol dir "que hagin de començar després que el programa s'acabi, o ni tan sols després que hagi començat: les millors avaluacions *ex-post* es dissenyen i es comencen a implementar *ex-ante*" (Ravallion, 2006). Entre les mesures més importants a prendre *ex-ante* hi ha aconseguir que els registres administratius incorporin informació

rellevant per a usos d'avaluació i millorin la seva qualitat. Se sol dir que els problemes que no tenen solució no són problemes sinó condicionants. En aquest sentit, la manca de dades adequades per a l'avaluació en els registres administratius és un condicionant a curt termini i un problema a llarg termini.

Notes:

1 *BLASCO, J. Avaluació del disseny. Barcelona: Ivàlua, 2009. (Guies pràctiques sobre avaluació de polítiques públiques; 3)*

3. MÈTODES PER A L'AVALUACIÓ D'IMPACTE

La qüestió fonamental que planteja l'avaluació d'impacte és mesurar fins a quin punt l'aplicació d'una determinada política sobre un conjunt d'individus modifica un determinat *outcome* d'interès, com ara llur renda o salut, respecte d'allò que aquests mateixos individus haurien experimentat en absència de la política. El que complica l'avaluació d'impacte és que la situació en absència del programa, l'anomenat contrafactual, és quelcom que per definició resulta inobservable pel grup d'individus que reben el programa. Així doncs, com ja hem esmentat en l'apartat anterior, el gran repte metodològic que planteja l'avaluació d'impacte és com definir un grup d'individus que, a més de no participar o beneficiar-se del programa o política, constitueixi un contrafactual creïble, és a dir que el seu nivell d'*outcome* pugui considerar-se equivalent al que hauríem observat per als beneficiaris de la política si aquesta no els hagués estat aplicada.

Els mètodes que s'empren en l'avaluació d'impacte difereixen entre si segons el procediment utilitzat per definir el grup d'individus que actuen com a contrafactual:

- D'una banda, els anomenats **dissenys experimentals** són aquells en què, partint d'una població de potencials beneficiaris del programa o política, els individus hi acaben participant o no d'acord a un mecanisme d'assignació purament aleatori; els individus que no hi participen, l'anomenat grup de control, constitueixen el contrafactual en aquest tipus de disseny.

- D'altra banda, la resta de mètodes disponibles, que reben el nom de **dissenys quasiexperimentals**, comparteixen la característica que la participació en el programa per part dels individus no ve definida per un procediment aleatori: ja sigui perquè són els mateixos individus els qui trien si participar-hi o no, ja sigui perquè algun altre agent pren aquesta decisió, o per totes dues coses alhora. En els dissenys quasiexperimentals, el contrafactual es defineix a partir dels individus que no participen en el programa, que constitueixen el que s'anomena grup de comparació.

Els apartats següents constitueixen una introducció breu, de caràcter no tècnic, als principals mètodes que es poden fer servir per establir l'impacte d'una política[1]. Començarem amb una introducció dels dos principals reptes als quals s'han d'enfrontar els diferents mètodes: maximitzar la robustesa amb què conclouen que el programa és la causa dels impactes observats (validesa interna) i la potencialitat per generalitzar les conclusions a altres programes, situacions i moments (validesa externa). A continuació, iniciarem l'exposició dels mètodes amb els experiments socials, ja que hi ha un ampli consens en el sentit que constitueixen el disseny més robust a l'hora d'avaluar l'impacte d'un programa. Per aquest motiu, tot i que són d'ús poc habitual, representen l'estàndard respecte al qual intenten emmirallar-se la resta de dissenys. La resta d'apartats consideren els diferents mètodes de caràcter

quasiexperimental més utilitzats: els anomenats dissenys abans-després, la tècnica de *matching* i el model de dobles diferències.

3.1. LA VALIDESA DE LES CONCLUSIONS

3.1.1. LA VALIDESA INTERNA

Els mètodes per a l'avaluació d'impacte que presentem en aquest capítol serveixen per **inferir una relació causal** entre una intervenció pública i determinats *outcomes*.

Utilitzem el concepte de **validesa interna** per referir-nos a la "veritat relativa" d'una inferència causal, és a dir, a la robustesa amb què es conclou que el programa és l'agent responsable dels impactes observats. La validesa interna no és una propietat de les metodologies sinó de les inferències concretes que es realitzen en cada avaluació, ja que un mateix mètode d'avaluació pot produir conclusions més o menys vàlides segons les circumstàncies i característiques del programa avaluat.

Les **amenaces a la validesa interna** són raons específiques per les quals és possible que estiguem parcialment o totalment equivocats a l'hora de fer una inferència causal. Concretament, són totes aquelles explicacions alternatives, a part del programa, que potencialment podrien ser responsables dels canvis observats en els *outcomes*. En cada avaluació, direm que el disseny metodològic és més o menys vàlid en tant que descarti convincentment aquestes explicacions alternatives. Les llistem a continuació de forma separada, tot i que algunes d'elles no són totalment independents:

1. **La història / factors contemporanis.** Fa referència a tots els esdeveniments que ocorren durant la implementació del programa i que poden tenir una influència sobre els *outcomes*. En l'exemple del programa d'atenció a les persones sense sostre, les variacions en el mercat de treball, la posada en marxa d'un programa de salut mental i els canvis en el control de la immigració formaven part de la història del programa, ja que s'esdevenien al mateix temps que el programa i tenien una influència sobre el nombre de persones que pernocten al carrer (la mesura de l'*outcome*), per la qual cosa podrien ser parcialment responsables dels canvis observats i, per tant, ser confosos amb l'impacte del programa. Se sol adreçar amb la identificació d'un grup de comparació que estigui exposat a esdeveniments externs iguals o similars.

2. **El biaix de selecció.** Una qüestió crítica quan s'identifica un grup de comparació és que sigui equivalent al grup de participants en totes les característiques que estan associades amb els *outcomes*, excepte pel fet que uns participen en el programa i els altres no.

El biaix de selecció es produeix quan aquesta assumpció no es compleix i existeix, des d'abans del programa, alguna diferència significativa entre els participants i el grup de comparació que pot ser potencialment responsable de les diferències observades al final del programa entre els *outcomes* d'uns i altres. Imaginem, per exemple, un programa de reforç lingüístic en català en què es proporciona formació en llengua catalana només als immigrants nouvinguts que ho sol·licitin, amb l'objectiu final de facilitar-los la inserció laboral. És molt possible que els qui s'hi apunten difereixin dels qui no ho han fet en característiques rellevants per a la inserció laboral: que el seu nivell educatiu sigui superior, que dominin més la llengua castellana o que tinguin més motivació per trobar una feina. És probable que, en absència del programa, els participants ho haguessin tingut igualment més fàcil per accedir al mercat laboral que els no participants. Per tant, si comparem l'evolució de la participació en el mercat laboral d'uns i altres és possible que part de la diferència en els *outcomes* es degui, en realitat, a aquestes diferències inicials en les seves característiques. L'amenaça del biaix de selecció és omnipresent en tots els dissenys no experimentals i adreçar-lo adequadament és, amb diferència, el principal repte metodològic de l'avaluació d'impacte.

3. **El desgast diferencial de la mostra (attrition).** Es tracta d'una forma del biaix de selecció que es produeix un cop iniciada l'avaluació. És relativament habitual que, al llarg de l'avaluació, alguns participants i membres del grup de comparació abandonin el programa, es neguin a seguir responent qüestionaris o simplement desapareguin. Aquestes pèrdues poden arribar a canviar la composició dels dos grups de manera que és molt possible que un i altre grup acabin sent diferents en alguna característica que estigui relacionada amb els *outcomes*, per més que inicialment estiguessin equilibrats. Aquesta diferència de composició entre un i altre grup pot ser la responsable dels canvis observats en els *outcomes*, que per tant, poden ser confosos amb l'impacte del programa. Suposem, per exemple, que en un programa destinat a prevenir recaigudes en exalcohòlics, aquells que millor es troben i més segurs estan de no recaure tendeixin a abandonar-lo abans de la seva finalització perquè el consideren innecessari, i se'ls perd la pista. En aquest cas, el grup de participants acaba estant compost per aquells amb un major risc de recaiguda mentre que el grup de comparació continua constituït per una barreja de persones amb riscos alts i baixos. En conseqüència, igual que passava amb el biaix de selecció, és possible que part de la diferència en els *outcomes* entre els dos grups es degui, en realitat, a aquestes diferències finals en la seva composició.

4. **Regressió a la mitjana.** És la tendència estadística que tenen els resultats extrems que es produeixen en un determinat moment de mesura dels *outcomes* a acostar-se a la mitjana de la població quan se'ls torna a mesurar un temps després. Això és perquè molts fenòmens impliquen una certa variació aleatòria: per exemple, un cap de setmana amb molts accidents de trànsit sol ser seguit per un altre amb un nombre menor, encara que les circumstàncies que determinen la propensió als accidents (el clima, el volum de

trànsit, etc.) no hagin variat, igual que les persones que van a psicoteràpia perquè estan molt estressades és probable que la següent vegada que hi vagin ho estiguin menys, fins i tot encara que no hagin rebut tractament. En general, aquesta amenaça s'ha de tenir en compte si la selecció per participar al programa es produeix precisament perquè la mesura de l'*outcome* ha estat substancialment alta o baixa. En aquestes situacions, és molt probable que en la següent mesura l'*outcome* millori per efecte de la regressió a la mitjana, i que aquest efecte es confongui fàcilment amb un efecte del programa.

5. **Efectes dels tests.** Algunes avaluacions consisteixen a realitzar un test a participants i membres del grup de comparació abans del programa (pretest) i després (posttest) a fi de poder estimar quin ha estat l'impacte de la intervenció. Ara bé, fer el pretest pot ensenyar a les persones a fer-ho millor en el test següent, o bé pot induir altres formes de reacció que es poden confondre amb els impactes del programa. Per exemple, si el test consisteix a fer proves de colesterol, pot ser que les persones cuidin més la seva dieta perquè saben que els el tornaran a mesurar. Igualment, en una prova de vocabulari, és possible que la gent amb mals resultats es prepari per a la següent perquè els fa vergonya tornar a fer-ho malament, o que senzillament ho faci millor perquè ja coneixen en què consisteix la prova i ja hi tenen una certa pràctica.

6. **L'efecte Hawthorne.** És un increment de l'*outcome* que experimenten les persones pel sol fet que algú té una atenció especial cap a ells, i no tant per l'efecte del programa en si mateix. Aquest efecte deu el seu nom a una sèrie d'estudis realitzats entre els anys 1927 i 1932 que van observar que els treballadors d'una planta elèctrica augmentaven la seva productivitat quan tenien la sensació que la direcció es preocupava d'ells, amb independència de la forma que prengués aquesta atenció. Així, tant abaixar la intensitat de la llum com apujar-la va provocar els mateixos impactes positius.

7. **Maduració.** El canvi natural o el creixement degut al mer pas del temps poden explicar les diferències entre els *outcomes* mesurats abans i després d'un programa. Per exemple, la millora de les capacitats cognitives dels infants, el temperament dels comportaments de risc dels adolescents o l'empitjorament de l'autonomia personal de la gent gran són fenòmens que es produiran entre el pretest i el posttest per efecte de la maduració i que es poden confondre amb els impactes del programa. Per adreçar aquesta amenaça és necessari disposar d'un grup de comparació de la mateixa edat perquè el fenomen de maduració afecti de manera similar a tots dos grups.

8. **Efectes dels instruments.** Si es produeix un canvi en l'instrument emprat per mesurar els *outcomes* en el pretest i el posttest, la variació en els *outcomes* pot reflectir els efectes d'aquest canvi tècnic en el sistema de recollida de dades i es pot confondre fàcilment amb els impactes del programa. És una amenaça freqüent quan l'avaluació fa un seguiment de sèries temporals llargues o quan la mesura depèn d'una valoració relativament subjec-

tiva que pot anar canviant al llarg del temps com, per exemple, l'apreciació del grau de desestructuració d'una persona sense sostre en el moment d'entrar al sistema.

9. **Externalitats (*spillovers*).** Es produeixen quan els no participants poden absorbir els beneficis del programa de forma indirecta, sovint pel fet d'estar en contacte amb els participants. Per exemple, en un programa pilot d'informació sexual a adolescents és possible que el grup de comparació millori els seus *outcomes* perquè els participants els han explicat allò que han après. Aquesta amenaça porta a la subestimació de l'impacte del programa.

3.1.2. LA VALIDESA EXTERNA

La validesa externa fa referència al grau en què les conclusions d'una avaluació poden ser generalitzades a altres programes similars, moments o llocs més enllà dels propis de la mateixa avaluació. Per exemple, si una avaluació demostra que una intervenció per al foment de l'emprenedoria empresarial ha estat efectiva en un estat de tradició industrial, als EUA, l'any 2006, podem concloure que també ho serà ara i a Catalunya?

De la mateixa manera que el disseny metodològic de l'avaluació determina el grau de validesa interna de les conclusions, també ho fa amb la validesa externa. En general, com més artificials i controlades siguin les condicions del programa per facilitar l'avaluació, menys plausible resulta pensar que aquestes condicions es reproduiran en un programa similar que no estigui subjecte a l'avaluació, i menys generalitzables seran les conclusions. Sovint, doncs, l'elecció d'un disseny metodològic ha de trobar un equilibri adequat entre la validesa interna i l'externa.

3.2. EXPERIMENTS SOCIALS

3.2.1. QUÈ SÓN I QUÈ ELS FA ROBUSTOS

Avaluar l'impacte d'una política pública mitjançant un experiment social és, des d'una perspectiva metodològica, molt similar a aplicar la lògica que segueixen els assajos clínics. Així, després de seleccionar un conjunt d'individus susceptibles de beneficiar-se dels potencials efectes positius de la política, se'ls assigna mitjançant un procediment aleatori, amb el seu consentiment, a un dels dos grups següents: d'una banda, l'anomenat grup de tractament, en què els subjectes participaran o rebran durant un cert període de temps la intervenció que caracteritza la política objecte d'avaluació (p. ex., un incentiu fiscal, un nou tipus de servei, etc.); de l'altra, l'anomenat grup de control, on els individus no rebran la

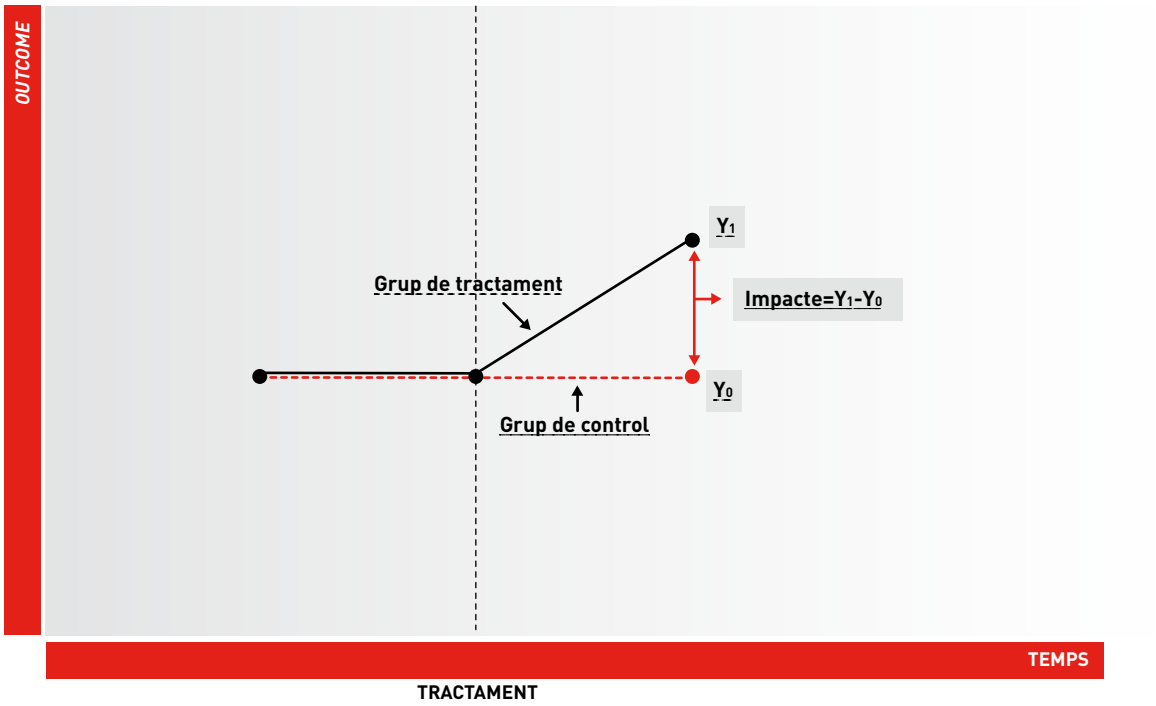
intervenció en qüestió. El fet de no rebre la intervenció no implica necessàriament que les persones que conformen el grup de control no rebin cap mena de tractament: n'hi ha prou que allò que rebin sigui diferent del que preveu la política que estiguem avaluant[2].

L'impacte de la política, en aquest tipus de disseny experimental, és molt fàcil de mesurar: només cal comparar, passat un cert temps, la mitjana que pren l'*outcome* d'interès (p. ex., trobar feina) entre els individus que formen part del grup de tractament i els que integren el grup de control (gràfic 4). Si aquesta diferència de mitjanes resulta estadísticament significativa, podrem concloure que la política té un efecte (positiu o negatiu) sobre l'*outcome* que estiguem analitzant[3].

Què és el que explica que, malgrat la seva senzillesa, els experiments socials constitueixin el disseny més robust a l'hora de mesurar l'impacte d'una política pública? El motiu cal buscar-lo en l'assignació aleatòria, que aconsegueix que els individus del grup de tractament i de control siguin equivalents en tots els factors que poden influir sobre l'*outcome* d'interès, amb l'excepció d'un de sol: la participació o recepció de la política que estiguem analitzant. A més, el fet que alguns d'aquests factors siguin inobservables per l'analista o de difícil mesura, com ara la motivació o l'interès dels individus, resulta del tot irrellevant: l'aleatorització permet que també els factors inobservables es distribueixin de manera similar entre tots dos grups de persones. Així doncs, donat que ambdós grups resulten equivalents en totes aquelles variables (observables o no) que poden influir sobre l'*outcome* d'interès, resulta legítim atribuir *causalment* qualsevol diferència en aquesta darrera variable a allò que distingeix els grups entre si: haver rebut o no la política.

En definitiva, la robustesa d'un experiment social com a mètode per avaluar impactes es deriva del fet que queda eliminada, per construcció, la principal amenaça a la validesa interna de qualsevol disseny d'avaluació: la possible existència d'un biaix de selecció (vegeu l'apartat 3.1.1). Aquesta gran virtut dels experiments socials, no obstant això, tan sols s'acabarà produint si l'equivalència entre el grup de tractament i de control es manté durant tot el període de temps que dura l'experiment. En aquest sentit, tal com s'explica en el proper apartat, hi ha diverses circumstàncies que poden aparèixer durant la fase d'implementació de l'experiment que erosionin llur validesa tant interna com externa.

Gràfic 4. Il·lustració d'un experiment social.



Font: Elaboració pròpia.

3.2.2. LÍMITS A LA VALIDESA DELS EXPERIMENTS

VALIDESA INTERNA

- **Fracàs de l'aleatorització.** La primera amenaça a la validesa interna d'un experiment social és que el procés d'assignació aleatori de casos i controls no hagi funcionat. La manera de comprovar aquest extrem passa, simplement, per contrastar si existeixen diferències estadísticament significatives entre tots dos grups en les mitjanes de totes aquelles variables (observables) que poden influir sobre l'*outcome* (per exemple, en cas d'un programa d'inserció laboral, aquestes variables podrien ser l'edat, el sexe, el nivell educatiu, etc.). És important comprovar-ho perquè, en cas que existeixin diferències, podrien portar a un biaix de selecció.

- **Biaix de selecció en les mostres (Sample Selection Bias).** Un problema amb què poden trobar-se els experiments socials és que, malgrat que hagin estat assignats aleatòriament als grups de tractament i control, alguns del individus del primer grup acabin no seguint el protocol de tractament (p. ex., no assisteixen als cursos de formació que preveu el programa), i/o alguns del grup de control hi acabin tenint accés. El risc que es produeixi aquest tipus de situacions depèn de la naturalesa de la intervenció que s'estigui analitzant: així, tot i que resulta plausible pensar que alguns tractats decideixin

no assistir a cursos de formació, sembla poc probable que aquest rebuig es produeixi si el tipus d'intervenció consisteix a rebre una transferència monetària. D'altra banda, pel que fa a la possibilitat que persones "control" acabin rebent la intervenció, l'aspecte clau a tenir en compte és la capacitat que puguin tenir els responsables de l'experiment per monitorar l'activitat dels gestors del programa i evitar situacions anòmales.

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

- **Desgast diferencial de la mostra.** En qualsevol experiment social hi ha un lapse de temps entre el moment de l'assignació aleatòria dels individus als grups de tractament i control, i el moment en què es mesura l'*outcome* d'interès per tal de valorar l'impacte de la política; si durant aquest lapse de temps hi ha individus del grup de tractament i/o de control que abandonen l'experiment, de tal manera que resulta impossible mesurar-ne els *outcomes*, direm que s'ha produït un fenomen de desgast mostral. Aquest desgast pot provocar un biaix en l'estimació de l'impacte si existeixen diferències en les característiques d'aquells que abandonen respecte dels que romanen, ja que desapareix l'equivalència entre els individus del grup de control i de tractament que s'havia aconseguit en el moment de l'aleatorització. En qualsevol cas, en aquelles situacions en què es produeix un desgast mostral que pot amenaçar la validesa dels resultats, hi ha tècniques estadístiques que permeten corregir (parcialment) el possible biaix resultant.

La naturalesa prospectiva dels experiments socials fa que les fases de planificació i disseny de l'avaluació siguin de crucial importància. El desgast mostral, l'existència d'externalitats i qualsevol altre factor que pugui esbiaixar els resultats de l'avaluació, han de ser anticipats i incorporats al disseny de l'experiment per tal d'eliminar-los o minimitzar-ne el seu abast. En cas contrari, quan l'experiment ja es troba en marxa, resulta pràcticament impossible refer-ne el disseny.

VALIDESA EXTERNA

En el cas d'un disseny experimental, la validesa externa dels resultats obtinguts (això és, la possibilitat d'extrapolar-los) es pot veure afectada per dos motius principals. En primer lloc, pot ser que la mostra d'individus a partir de la qual s'hagin definit els grups de trac-

tament i de control no sigui representativa de la població a la qual pretenem extrapolar els resultats; aquest seria el cas, per exemple, d'un experiment social que s'hagués portat a terme en una determinada comarca de Catalunya que no fos representativa de la població catalana. D'altra banda, també pot passar que el mateix programa no resulti representatiu, és a dir, que la manera en què aquest opera en condicions experimentals no pugui reproduir-se a una escala superior (per exemple, en el cas d'un programa de reforç educatiu, pot passar que el nivell de motivació dels professionals no sigui el mateix, o que la insuficiència de recursos dilueixi alguns elements del programa quan s'aplica a gran escala, etc.).

3.2.3. PER QUÈ NO HI HA MÉS EXPERIMENTS SOCIALS?

En les darreres dècades, els experiments socials han tingut un creixement notable, sobretot als països anglosaxons i també en alguns països en vies de desenvolupament. Als EUA, país capdavanter en aquest sentit, s'han fet experiments socials per avaluar canvis en les polítiques educatives (Krueger, 1999), reformes en els programes de manteniment de rendes per a persones pobres (Moffit, 2004), o també experiències innovadores en el camp de les polítiques actives d'ocupació (Heckman, 1997). També hi hagut experiments socials en altres països del continent americà, com ara Mèxic, on l'avaluació experimental del programa PROGRESA va tenir un impacte notable a nivell internacional (Skoufias, 2005). A Colòmbia, Xile o l'Argentina, entre altres, també han avaluat mitjançant experiments diverses polítiques en àmbits laborals, educatius o dels serveis socials, així com en altres països dels continents asiàtic i africà que reben fons provinents de l'ajuda internacional[4]. El quadre 6 il·lustra les característiques d'aquest tipus de dissenys de la mà d'un experiment social concret: l'avaluació d'una reforma organitzativa dels serveis socials i sanitaris per a persones grans al Quebec (Béland [et al.], 2006).

En qualsevol cas, com que la majoria d'analistes considera els experiments socials el disseny més robust per avaluar l'impacte d'una política (el *gold standard*), resulta fins a cert punt paradoxal que no hi hagi molts més experiments socials dels que hi ha.

Un primer factor a considerar és l'elevat cost que, en general, tenen aquest tipus de dissenys: d'una banda, atès que la minimització de les amenaces a la validesa de l'experiment requereix un rigorós procés de planificació *ex-ante*, la negociació entre les diverses parts implicades sobre aquestes qüestions pot resultar força important en termes de temps; d'altra banda, si la hipòtesi és que els efectes de la política no siguin immediats i es pretén poder extrapolar els resultats de l'experiment a d'altres àrees del país, caldrà treballar amb mostres de controls i tractaments de grandària suficient (milers de persones) que caldrà seguir durant un ampli període de temps[5].

En qualsevol cas, més enllà de les consideracions econòmiques, l'argument habitual que utilitzen els qui s'oposen als experiment socials té un rerefons ètic: resulta inadequat privar determinats individus (els del grup de control) dels beneficis que suposa una nova política utilitzant un mecanisme tant arbitrari com l'aleatorització. La rèplica per part d'aquells que veuen en els experiments socials una eina adequada d'avaluació se sustenta en tres consideracions. La primera és que la presumpció que s'està privant alguns individus de quelcom de beneficiós 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 mena 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 l'aleatorització pot considerar-se 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 aleatoritzat (*randomized phase-in*): allò que s'aleatoritza és el moment del temps en què diferents grups d'individus començaran a rebre el nou programa.

En qualsevol cas, més enllà de quines siguin les raons que hi ha darrere de l'escassetat d'experiments socials, el cert és que moltes de les avaluacions d'impacte que es porten a terme arreu fan servir dissenys de caràcter no experimental. Dedicarem els propers apartats a descriure breument els principals mètodes disponibles a tal efecte.

QUADRE 7
EXEMPLE D'EXPERIMENT SOCIAL: SISTEMA INTEGRAT D'ATENCIÓ SANITÀRIA DEL QUEBEC

CONTEXT: En molts casos, la manca d'autonomia de les persones grans ve motivada pel patiment de malalties cròniques i, per això, les necessitats d'atenció d'aquest col·lectiu són tant sanitàries com socials. No obstant això, a la majoria de països desenvolupats, inclòs el Canadà, la resposta assistencial que proporcionen els sistemes sanitari i social acostuma a portar-se a terme sense cap mena de coordinació.

OBJECTIU: L'equip investigador pretenia avaluar en quina mesura un sistema integrat d'atenció, anomenat SIPA per les seves sigles en francès, permetria millorar la salut de les persones grans dependents del Quebec, augmentar la satisfacció dels seus cuidadors i reduir els costos assistencials totals.

TIPUS D'ESTUDI I INTERVENCIÓ: L'avaluació del nou model integrat d'atenció es va portar a terme mitjançant un experiment aleatoritzat amb grup de control. Els pacients assignats al *grup de tractament* (606) van passar a ser atesos per equips multidisciplinars que no tan sols proporcionaven directament els serveis comunitaris socials i sanitaris (atenció domiciliària, centre de dia, centre de salut, infermeria domiciliària, etc.), sinó que també coordinaven l'atenció hospitalària i la institucionalització social (residències d'assistits) dels pacients. D'altra banda, els individus del *grup de control* (624) van continuar rebent l'atenció de la manera habitual, o sigui, mitjançant l'acció independent dels sistemes sanitari i social del Quebec.

***OUTCOMES*:** Durant 22 mesos, es va obtenir informació de registre sobre els serveis sanitaris i socials utilitzats pels pacients assignats a tots dos grups, incloent-hi també els costos de l'atenció rebuda en cada cas. Addicionalment, en el moment de començar l'estudi i transcorreguts 12 mesos, es va utilitzar una enquesta per obtenir informació sobre l'estat de salut de la persona gran, la satisfacció i la càrrega suportada pels cuidadors, així com les despeses privades assumides per la família en relació amb la cura de la persona dependent.

RESULTATS: Els pacients atesos mitjançant el model SIPA van fer una major utilització dels serveis sanitaris i socials de caràcter comunitari, però la seva probabilitat de patir episodis d'hospitalització innecessariament llargs (*bedblocking*) fou menor que la de les persones del grup de control. No obstant això, pel que fa a la resta de serveis sanitaris i socials considerats, no es va detectar cap mena de diferència entre ambdós grups: van utilitzar les urgències hospitalàries amb la mateixa intensitat, van ser ingressats als hospitals amb la mateixa freqüència i van tenir la mateixa probabilitat d'acabar ingressats en una residència d' assistits.

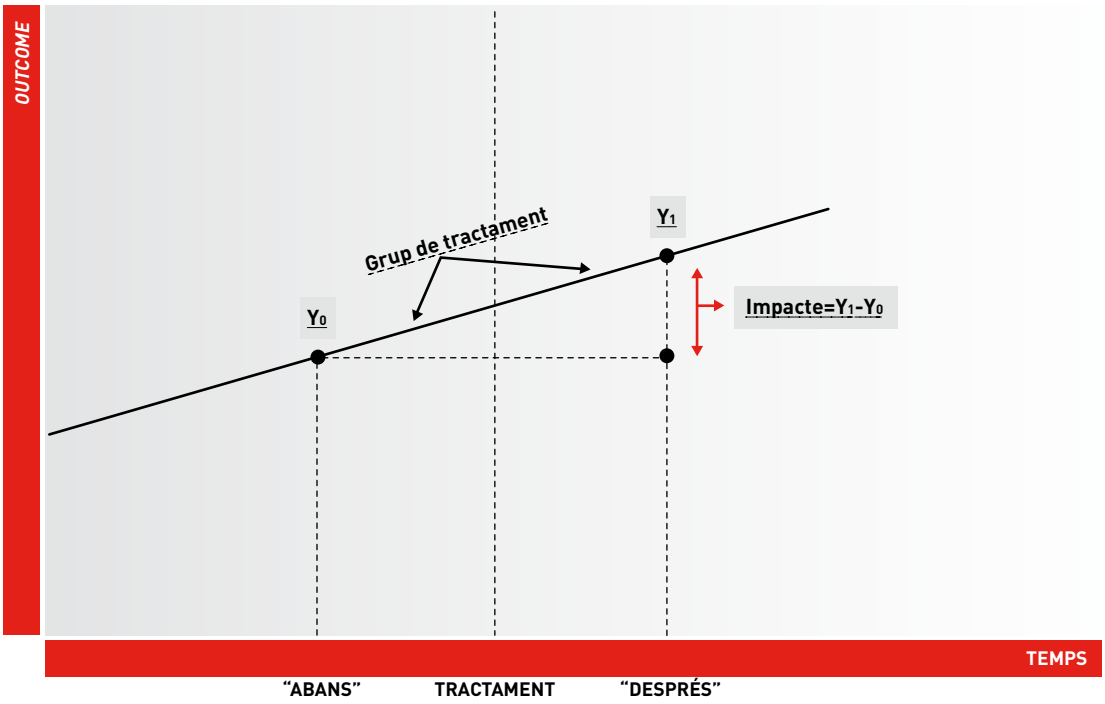
En termes de costos, si bé els pacients del model SIPA van tenir unes despeses mitjanes inferiors en els serveis que impliquen la institucionalització dels individus (hospitals i residències), aquest efecte es va veure totalment compensat per un increment en la despesa mitjana dels serveis comunitaris, de tal manera que el cost total mitjà d'ambdós grups va acabar sent el mateix. D'altra banda, tot i que la satisfacció dels cuidadors informals dels "pacients SIPA" va augmentar, no es van detectar diferències significatives pel que fa a la "càrrega" suportada. Finalment, tampoc no hi va haver diferències entre ambdós grups pel que fa a l'evolució de l'estat de salut dels pacients tractats en cada cas.

Font: Elaboració pròpia a partir de Béland [et al.] (2006).

3.3. DISSENYS SENSE GRUP DE CONTROL: ABANS-DESPRÉS I SÈRIES TEMPORALS

El mètode quasiexperimental més simple per avaluar impactes i, com veurem, també el menys robust, és l'anomenat disseny abans-després. La seva aplicació requereix disposar d'informació relativa als beneficiaris de la política tant abans com després de la seva posada en marxa. L'impacte de la política s'obté, simplement, calculant la diferència entre la mitjana de l'*outcome* per la mostra de beneficiaris en cadascun dels dos moments esmentats. El contrafactual es defineix reflexivament, d'aquí que aquest disseny es conegui també amb el nom de controls reflexius, en el sentit que la mesura d' "allò que hauria passat als beneficiaris en absència de la política" s'obté a partir de l'experiència d'aquests mateixos individus abans que la política existís (gràfic 5).

Gràfic 5. Il·lustració d'un disseny abans-després.



Font: Elaboració pròpia.

El supòsit clau perquè aquest mètode estimi correctament l'impacte d'una política és que no hi pot haver cap altre factor, diferent del programa, que hagi pogut afectar l'*outcome* d'interès entre els dos moments de recollida de dades. En la majoria de casos, però, resulta evident que la plausibilitat d'aquest supòsit serà mínima. Imaginem, a tall d'exemple, una hipotètica reforma que doni més autonomia de gestió als centres amb l'objectiu de reduir les taxes de fracàs escolar. En aquest cas, si ens aproximem a la mesura de l'impacte mitjançant un disseny abans-després, els possibles canvis que observem en l'evolució de les taxes de fracàs escolar poden haver estat provocats per múltiples factors diferents de la reforma: un decrement de les ràtios alumnes/professor fruït de l'evolució demogràfica, una reforma curricular, canvis en el perfil sociodemogràfic dels pares, etc.

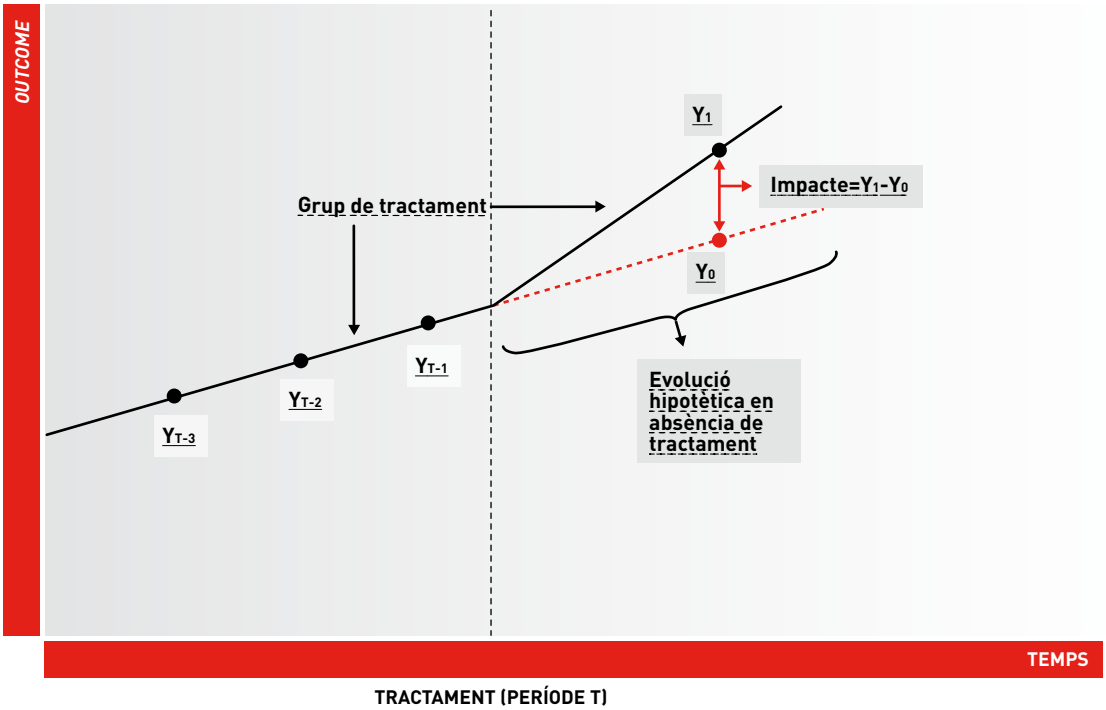
Però, a més de les amenaces a la validesa interna provocades pel que en l'apartat 3.1 anomenàvem "història o factors contemporanis", els dissenys abans-després són també molt vulnerables a amenaces a la validesa interna, especialment els anomenats fenòmens de maduració i regressió a la mitjana. En essència, com que aquest tipus de disseny es troba mancat d'un grup de comparació genuí sobre el qual construir un contrafactual creïble, sempre hi ha el dubte que les variacions observades en l'*outcome* al llarg del temps no s'haguessin produït de totes maneres, encara que la política avaluada no hagués tingut lloc.

Així doncs, malgrat que s'utilitzen amb força profusió, els dissenys abans-després són un mètode molt poc robust. És per això que, sempre que sigui possible, optarem per altres mètodes que basin la seva estratègia d'identificació en la comparació de grups de persones beneficiàries i no beneficiàries de la política. Què fer quan resulta totalment impossible construir un grup de comparació no beneficiari de la política, com és típicament el cas quan una política s'introdueix en tot el territori i afecta tota la població? En aquestes circumstàncies, només si estem molt segurs que els impactes esperats de la política són força immediats i que no hi ha factors contemporanis que influeixin sobre l'*outcome*, podríem arribar a considerar un disseny abans-després; en canvi, si aquestes circumstàncies no es donen, caldria reconsiderar seriosament la conveniència de portar a terme una avaluació d'impacte quantitativa.

Els anomenats models de **sèries temporals interrompudes** constitueixen l'altre gran tipus de disseny quasiexperimental que, igual que els dissenys abans-després, miren d'estimar l'impacte d'una política sense utilitzar un grup de comparació. En certa mesura, constitueixen una variant refinada dels dissenys abans-després, ja que la seva principal característica és que utilitzen informació sobre múltiples períodes de temps anteriors i posteriors a la introducció de la política que es pretén avaluar. Així doncs, en comparació amb un model abans-després, el contrafactual reflexiu d'aquest tipus de dissenys resulta més creïble ja que disposem de més informació per estimar què hauria passat en absència de la política.

L'estratègia d'identificació dels impactes que utilitzen els dissenys de sèries temporals interrompudes és senzilla. A partir de les observacions disponibles sobre l'evolució de l'*outcome* abans de la intervenció, s'utilitzen tècniques estadístiques per mirar de modelitzar-ne el seu comportament "normal" en absència de la intervenció, tot tenint en compte la possible influència que hagin pogut tenir altres factors. A continuació, aquest comportament normal es projecta als períodes posteriors a la introducció de la política, i es contrasta fins a quin punt existeixen discrepàncies entre les prediccions del model i els valors realment observats; si hi ha, s'atribueixen aquestes discrepàncies a l'existència de la política (gràfic 6). No obstant això, tot i que la idea subjacent és simple, val a dir que els models de sèries temporals són tècnicament complexos i llur aplicació exigeix coneixements avançats d'estadística.

Gràfic 6. Il·lustració d'un disseny de sèries temporals interrompudes.



Font: Elaboració pròpia.

En qualsevol cas, tot i que els models de sèries temporals constitueixen un mètode més robust que els dissenys abans-després per avaluar impactes, cal tenir present que el contrafactual se segueix construint de forma reflexiva. Així doncs, tot i que puguem tenir en compte la influència d'altres factors sobre l'evolució de l'*outcome*, aquesta influència es modelitza en funció d'una informació que no pertany al mateix període de temps en què realment opera la política. En definitiva, tot i que suavitzats, segueixen resultant d'aplicació el mateix tipus de cauteles que plantejàvem en el cas dels dissenys abans-després, o sigui, limitar el seu ús a situacions en què, d'una banda, no resulti possible construir un grup de comparació i, de l'altra, hi hagi una quantitat molt reduïda d'explicacions alternatives sobre el perquè de l'evolució de l'*outcome* després de la introducció de la política. En la resta de casos, resulta recomanable explorar altres tipus de dissenys, com els que s'expliquen tot seguit, en què l'impacte es mesura utilitzant persones no beneficiaries (grup de comparació) contemporànies d'aquelles que es beneficien de la política (grup de tractament).

3.4. LA TÈCNICA DEL *MATCHING*

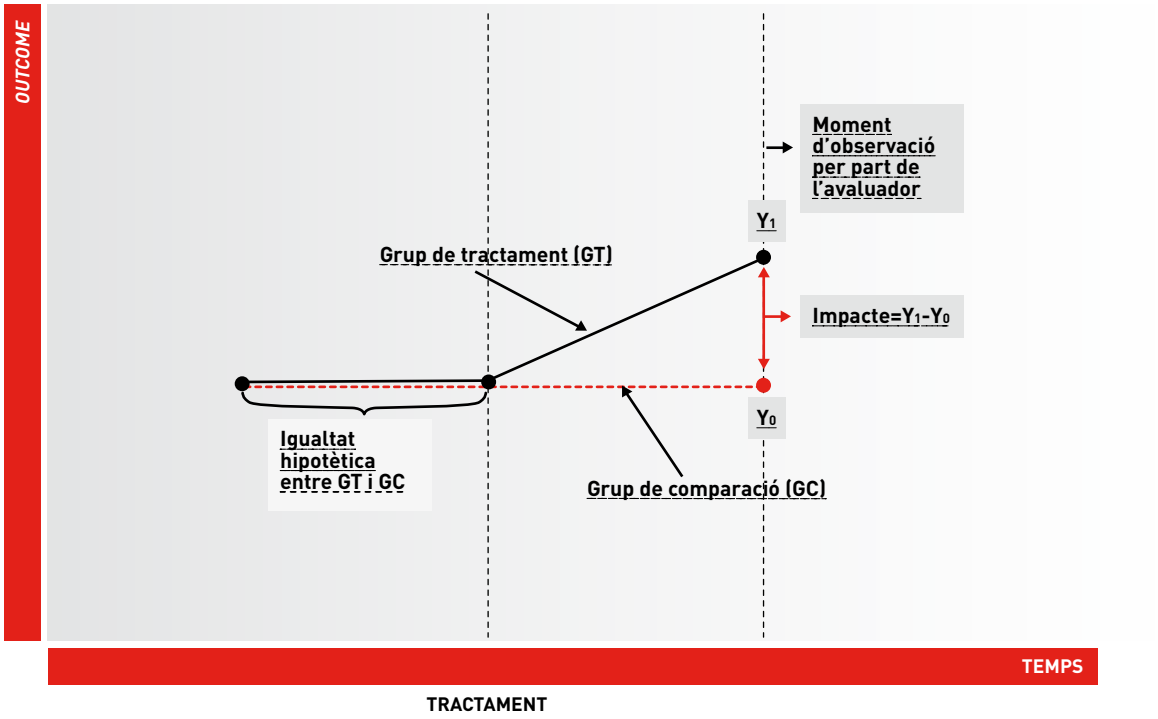
3.4.1. QUÈ ÉS?

Aquesta tècnica imita un experiment amb assignació aleatòria de tractament mitjançant la creació d'un grup de control *ex-post* que s'assembla al màxim possible al grup de tractament per a aquelles característiques rellevants observables. L'aplicació d'aquest mètode per avaluar l'impacte d'una política pot considerar-se en aquells casos en què, amb posterioritat a la intervenció pública, disposem d'informació tant d'una mostra d'individus que han estat beneficiaris del programa com d'una altra de persones que no ho han estat. En concret, per a cadascun dels individus d'ambdós grups, cal tenir informació sobre el valor que pren en cada cas l'*outcome* d'interès i també sobre tots aquells factors (característiques dels individus, entorn en el qual viuen, etc.) que, d'una banda, poden haver determinat el procés pel qual els individus han decidit participar en el programa i, de l'altra, poden tenir efectes sobre el valor que pren l'*outcome* d'interès.

El que el mètode de *matching* proposa és utilitzar tota la informació anterior per construir un grup de comparació entre els individus que no es beneficien del programa. Per fer-ho, el mètode busca, per a cadascun dels individus que composen la mostra de tractats, una parella o *match* (d'aquí el nom de la tècnica) que sigui el més semblant possible en el sentit que acabem de descriure.

La pretensió darrera de la tècnica del *matching* és obtenir, mitjançant procediments estadístics, allò que els experiments socials obtenen mitjançant l'assignació aleatòria, a saber, que el grup d'individus que ens ha de servir per construir el contrafactual sigui el més semblant possible al grup d'individus que reben el programa, a fi de minimitzar tant com es pugui el biaix de selecció. Però mentre que una assignació aleatòria veritable distribueix de forma *equitativa* les característiques observables i les no observables entre el grup control i el de tractament, el *matching* només distribueix *equitativament* les característiques observables. En altres paraules, assumeix que no hi ha cap variable rellevant no observable que difereixi sistemàticament entre el grup de tractament i el de comparació i que, per tant, l'*outcome* del grup de tractament si no hagués participat o s'hagués beneficiat del programa (és a dir, el contrafactual) equival a l'*outcome* del grup de comparació que, realment, no hi ha participat (gràfic 7).

Gràfic 7. Il·lustració d'un disseny basat en la tècnica del *matching*.



Font: Elaboració pròpia.

Un exemple pot ajudar-nos a acabar de comprendre la lògica d'aquest tipus de disseny. Imaginem que el Departament de Salut posa en marxa una política d'incentius destinada a incrementar la prescripció de genèrics per part dels metges d'atenció primària. Suposem que el percentatge de medicaments genèrics que prescriuen els facultatius només es troba influït per l'edat del metge i pel seu sexe, i que la decisió de participar o no en el programa d'incentius és voluntària. Així les coses, si tinguéssim la sort que no hi hagués diferències pel que fa al sexe i l'edat dels metges que hi participen i dels que no, una simple comparació de mitjanes entre ambdós grups pel que fa al percentatge de prescripció de genèrics ens proporcionaria una bona estimació de l'impacte de l'esquema d'incentius. I si, per contra, observem que la distribució per sexe i edat dels participants és diferent de la dels no participants? Llavors no podríem atribuir la diferència en la mitjana dels *outcomes* exclusivament a la intervenció, ja que estarà també motivada pel fet que ambdós col·lectius són diferents. En aquest cas, una possible estratègia seria construir el grup de comparació seleccionant únicament aquells metges no participants que garantissin un percentatge de dones i una distribució per edats idèntics als del grup de participants: així, per a cada dona participant d'entre 30 i 35 anys, buscaríem una dona d'igual edat en el grup de no participants. Una lògica molt similar a la que acabem de descriure és la que utilitza el *matching* per tal d'intentar obtenir estimacions no esbiaixades de l'impacte de les polítiques.

3.4.2. COM S'IMPLEMENTA: PROPENSITY SCORE I APARELLAMENT

L'exemple anterior és poc realista en el sentit que resulta evident que la prescripció de genèrics es troba determinada per més factors que el sexe i l'edat dels metges. En general, com que el nombre de variables susceptibles d'influir tant sobre la decisió de participar en un programa com sobre l'*outcome* d'interès serà força elevat, resulta impossible realitzar un aparellament com el que descrivíem en l'exemple anterior.

L'alternativa passa per reduir la dimensionalitat del problema i definir la major o menor similitud entre tractaments i controls a partir d'un sol nombre: l'anomenat *propensity score (PS)*. El PS mesura la probabilitat que un individu, donades les seves característiques, decideixi participar en el programa; aquesta probabilitat s'obté a partir d'un model d'elecció discreta, com ara un lògit o un pròbit[6].

El pas següent consisteix a realitzar els aparellaments entre participants i no participants basant-nos en el PS d'uns i altres. Hi ha diversos mètodes per definir com es constitueixen les parelles. El més senzill és el que s'anomena "el veí més proper" *(nearest-neighbour caliper)* i consisteix a formar tan sols aquelles parelles participant-no participant en què la diferència entre el PS d'un i altre sigui inferior a un cert nombre predeterminat. Aquest mètode permet que estiguin equilibrades la mostra de participants i la mostra final de no participants amb què els acabem comparant, si més no pel que fa a les variables que hem considerat a l'hora de modelitzar la participació en el programa.

El darrer pas consisteix a estimar l'impacte de la intervenció. En aquest sentit, igual que en el cas dels experiments socials, la tècnica de càlcul és ben senzilla: n'hi ha prou de computar la mitjana aritmètica de les diferències en *outcomes* de les diferents parelles construïdes, i verificar si aquesta mitjana és significativament diferent de zero o no.

3.4.3. LIMITACIONS

El supòsit bàsic que la tècnica del *matching* necessita per obtenir estimacions consistents de l'impacte d'una política és que, en mitjana, un cop s'ha tingut en compte l'efecte de les variables condicionants (el sexe, l'edat, l'especialitat, etc., en el cas de l'exemple dels metges), els participants haguessin obtingut el mateix *outcome* que els no participants si la política no hagués existit. O, dit d'una altra manera, el supòsit fonamental és que no existeix el que tècnicament s'anomena "selecció en variables no observables", és a dir, no hi ha cap factor que no hagi estat tingut en compte per l'analista que influeixi simultàniament sobre la probabilitat de participar en el programa i sobre l'*outcome* d'interès. En cas contrari, com que res no garanteix que l'aparellament hagi generat mostres de tractaments i controls equilibrades pel que fa a aquests factors no observats, la mesura de l'impacte que

obtinguem pot patir un biaix important respecte del seu autèntic valor. En aquest sentit, seguint amb l'exemple dels incentius als metges, aquest seria el cas si existissin diferències (no observables) de motivació entre participants i no participants.

Intuïtivament, per tal de minimitzar el risc que hi hagi un biaix de selecció en les pròpies estimacions, sembla obvi que el que hauria de fer l'analista és intentar aplicar la tècnica del *matching* utilitzant un conjunt de variables de control el més ampli possible; en concret, s'haurien de tenir en compte totes aquelles variables per a les quals existís evidència que influeixen tant la participació com l'*outcome* d'interès. En aquest sentit, si per a alguns d'aquests factors no existeix informació (és a dir, si aquests factors són inobservables), la credibilitat dels resultats obtinguts quedarà erosionada.

Amb la intenció d'il·lustrar les possibilitats que ofereix la tècnica del *matching* a la pràctica, el quadre següent conté la descripció d'una avaluació d'impacte que, seguint aquesta metodologia, va intentar esbrinar l'efectivitat dels principals programes de formació ocupacional existents a Catalunya.

QUADRE 8
AVALUACIÓ DE LA FORMACIÓ OCUPACIONAL A CATALUNYA

El Servei d'Ocupació de Catalunya (SOC) porta a terme un ampli conjunt d'accions formatives, dirigides a diversos col·lectius d'aturats, l'objectiu de les quals és millorar les possibilitats que aquestes persones trobin una feina i la mantinguin. Els programes en marxa comprenen, entre altres, els següents: *Pla FIP*, destinat prioritàriament a persones desocupades majors de 65 anys, aturats de llarga durada, discapacitats, etc.; *Centres d'Innovació i Formació Ocupacional (CIFO)*, cadascun d'ells especialitzat en una o vàries famílies professionals; *Igualtat d'Oportunitats*, programa de formació dirigit específicament a dones; etcètera. L'any 2008, per encàrrec del SOC, un equip d'investigadors dirigit pel professor Toharia va portar a terme una avaluació dels impactes d'aquests programes fent servir la tècnica del *matching* (Toharia [et al.], 2008). Els principals ingredients metodològics d'aquesta avaluació varen ser els següents:

- ***Outcomes*:** Situació laboral de la persona durant els 8 trimestres posteriors a l'any en què van tenir lloc els programes avaluats.

- **Grups de tractament i de control:** Es van definir 8 grups de tractament diferents, un per a cadascun dels 8 programes de formació ocupacional avaluats (Pla FIP, CIFO, Igualtat d'Oportunitats, etc.). Addicionalment, es van definir mitjançant la tècnica de *matching* 8 grups de comparació constituïts per demandants d'ocupació que no s'havien beneficiat de cap dels programes formatius del SOC, però que segons el seu *propensity score*, tenien característiques similars a les persones beneficiàries dels diferents programes.

- **Variables del *propensity score*:** Sexe, edat, nacionalitat, nivell d'estudis, àmbit de recerca, temps d'inscripció, alta nova, aturat de llarga durada, recepció de prestacions, nombre d'ocupacions demandades i província de residència.

Els resultats obtinguts assenyalen que tant els CIFO com el Pla FIP augmenten la probabilitat d'estar ocupats dels participants respecte dels no participants. En canvi, pel que fa al programa d'Igualtat d'Oportunitats i a les Accions Integrades (dirigides a persones amb dificultats especials), els impactes estimats sobre l'ocupació van ser nuls. Finalment, en el cas dels Programes de Garantia Social, dirigits a joves que finalitzen l'ESO sense acreditar-la, es va detectar un efecte negatiu sobre la probabilitat d'estar ocupat que tendeix, però, a disminuir ràpidament en el temps; ara bé, s'ha de tenir en compte que aquest programa és el de major durada i, per tant, cal pensar que els efectes tendeixen a produir-se a més llarg termini.

Font: Elaboració pròpia a partir de Toharia [et al.] (2008).

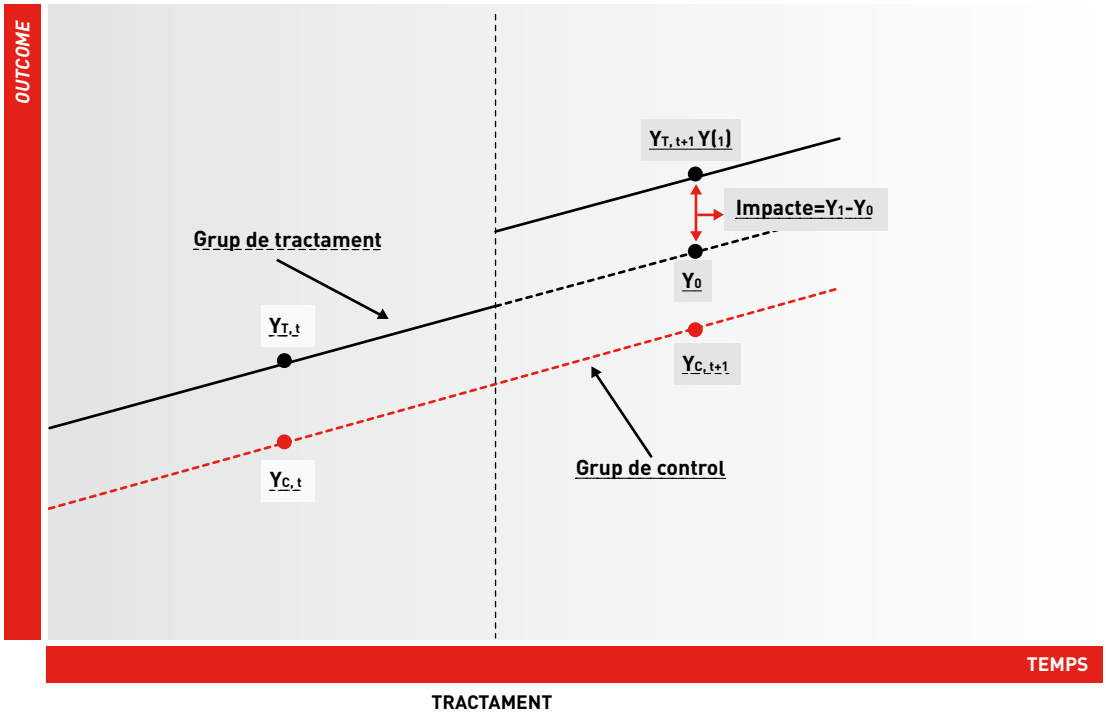
3.5. EL MODEL DE DOBLES DIFERÈNCIES

3.5.1. DEFINICIÓ I AVANTATGES

Aquesta tècnica s'aproxima a la quantificació de l'impacte d'un programa definint l'efecte no en termes de la diferència posttractament en el nivell de l'*outcome* per als beneficiaris i per als no beneficiaris, sinó com la diferència en la variació de l'*outcome* abans i després de la política en ambdós grups[7]. Així doncs, en definir l'impacte d'aquesta manera, la

tècnica de dobles diferències reconeix explícitament que part de la variació temporal en l'*outcome* d'aquells que reben la política s'hauria produït en qualsevol cas, i que la manera de mesurar-la és a través del canvi en l'*outcome* dels no beneficiaris durant el mateix període. La millor forma d'entendre aquesta tècnica és a través de la seva representació gràfica.

Gràfic 8. Il·lustració d'un model de dobles diferències.



Font: Elaboració pròpia.

Així, tal com pot observar-se a al gràfic 8, l'impacte que s'estima amb un model de dobles diferències és la diferència entre l'*outcome* dels beneficiaris (tractaments) després de la política ($Y_{T,\,t+1}$; el nostre Y_1) i el valor d'aquest *outcome* per a aquest mateix col·lectiu en absència del programa, el famós contrafactual, representat a la figura com a Y_0. L'essència del mètode és que aquest contrafactual s'obté projectant el nivell de l'*outcome* dels beneficiaris abans de la política ($Y_{T,\,t}$) a una determinada taxa de variació: l'observada pel que fa als *outcomes* dels "controls" entre el moment anterior ($Y_{C,\,t}$) i posterior a la introducció de la política ($Y_{C,\,t+1}$). En definitiva, és fàcil demostrar que la mesura de l'impacte pot expressar-se en termes analítics com la "diferència de diferències" (d'aquí el nom de la tècnica) següent:

$$(Y_{T,\,t+1} - Y_{T,\,t}) - (Y_{C,\,t+1} - Y_{C,\,t})$$

on $Y_{T,\,t}$ i $Y_{T,\,t+1}$ són les mitjanes de l'*outcome* per al grup de tractament abans i després de la política, i $Y_{C,\,t}$ i $Y_{C,\,t+1}$ les del grup de comparació.

El model de dobles diferències (DD), en la mesura en què fa servir informació d'abans i després de la posada en marxa de la política tant per als beneficiaris com per als no beneficiaris, és capaç de superar algunes de les limitacions que amenaçaven la validesa interna d'altres tipus de dissenys[8].

En primer lloc, si el comparem amb un disseny abans-després, la utilització per part del model DD d'un grup de comparació permet prevenir el possible biaix provocat per factors contemporanis a la política que poden tenir efectes sobre l'*outcome* d'interès. Per exemple, si estem interessats a avaluar l'efecte d'un programa de formació sobre les possibilitats de trobar feina dels aturats, factors d'aquest estil serien variacions en la taxa d'atur, modificacions en la normativa laboral, etc. Igualment, fruït de l'existència d'un grup de comparació, un model de DD que, analitzés l'impacte d'un programa de beques sobre el rendiment escolar seria també menys sensible que un disseny abans-després a patir biaixos per regressió a la mitjana (ja que aquest fenomen afectaria igual a beneficiaris i no beneficiaris).

D'altra banda, en la mesura en què el que s'estima és la diferència entre tractaments i controls en la variació dels *outcomes* i no la diferència en el nivell en si, els models DD poden eliminar algunes de les fonts del biaix de selecció que l'existència de factors inobservables provocava en el cas del *matching*. En concret, el tipus de factors inobservables que no tenen efecte sobre la consistència de la mesura d'impacte d'un model DD són aquells que no varien al llarg del temps. Podem il·lustrar aquesta propietat a partir de l'exemple esmentat sobre un hipotètic programa de formació. Suposem que la motivació fos un factor inobservable i que aquesta variable es distribuís de manera diferent entre els individus que participen en el programa (més motivats) i els que no ho fan (menys motivats). En aquest cas, és evident que part de la diferència en el nivell dels *outcomes* d'ambdós grups (tant abans com després de la intervenció) s'explicaria per la influència d'aquest factor; ara bé, com l'*impacte* que mesura el model DD no es dóna en termes de nivells sinó de taxes de variació de l'*outcome*, el fet que la diferència de motivació no variï al llarg del temps fa que aquest factor no pugui haver estat la causa de l'evolució diferencial de l'*outcome* en el grup de tractament respecte del de control.

3.5.2. LIMITACIONS

Els models de dobles diferències, malgrat els seus avantatges, no es troben exempts de veure amenaçada llur validesa interna si no es compleixen els dos supòsits que permeten a aquest tipus de disseny identificar correctament l'impacte d'una política pública.

El primer d'aquests supòsits és que tant els participants com els no participants han de reaccionar de la mateixa manera davant dels factors contemporanis a la política que, més

enllà d'aquesta, poden influir sobre l'*outcome* d'interès. En el cas del programa de formació abans esmentat això significa que, per exemple, si es produeix una millora en un factor que influeix sobre la probabilitat que els individus tenen de trobar feina, com ara una reducció en la taxa d'atur, el seu efecte sobre tractaments i controls ha de ser el mateix. En aquest cas, la violació d'aquest supòsit podria produir-se si l'augment de l'ocupació s'hagués concentrat en feines d'elevada qualificació, i els nivells formatius dels tractaments fossin superiors al dels controls, ja que llavors la millora induïda per la caiguda de l'atur seria superior entre els primers.

Hi ha dues formes d'intentar mitigar les possibles sospites que pugin existir sobre el compliment del supòsit "d'igualtat de reacció davant de factors contemporanis". En primer lloc, si existeix informació sobre múltiples períodes de temps previs a la introducció de la política, podem contrastar si efectivament els *outcomes* de tractaments i controls han evolucionat de manera similar quan s'han produït variacions en determinats factors que també tenen influència sobre l'*outcome* (la taxa d'atur, en el nostre exemple). L'altra possibilitat que podem aplicar quan no existeix informació retrospectiva és estimar el model DD després d'haver seleccionat els grups de tractament i control utilitzant la tècnica del *matching*. D'aquesta manera, com que el *matching* ens garanteix una elevada similitud entre els dos grups, cal pensar que les possibilitats que uns i altres reaccionin de la mateixa manera davant de factors contemporanis augmenta i, per tant, també la consistència dels resultats del model DD.

El segon supòsit que s'ha de satisfer perquè el model DD proporcioni estimacions no esbiaixades de l'impacte d'una política és que no poden existir diferències entre tractaments i controls en característiques no observables que variïn al llarg del temps. Si hi són, el fet que els models DD mesurin l'impacte com la diferència entre tractaments i controls en la variació de l'*outcome* no permet en aquest cas eliminar possibles biaixos. Així, seguint amb l'exemple del programa de formació, si la motivació de tractaments i controls varia al llarg del temps, i no podem observar aquesta variable, no podrem estar plenament segurs que aquest factor no sigui la causa de l'evolució diferencial de l'*outcome* en el grup de tractament respecte del de control i, per tant, que la magnitud de l'impacte estimat per a la política no en sobreestimi el seu efecte real. Així doncs, si volem que els resultats d'una avaluació d'impacte que utilitzi un disseny DD resultin creïbles, haurem d'oferir arguments que permetin descartar l'existència de característiques inobservables que varien en el temps de forma diferent entre tractaments i controls.

El següent quadre il·lustra les possibilitats dels models DD a través d'una aplicació portada a terme en el nostre entorn. En concret, el cas comentat és el d'una avaluació d'impacte que estima, mitjançant un model DD, els efectes que podrien derivar-se d'una major cobertura per part del sector públic de l'atenció bucodental dels infants.

QUADRE 9
AVALUACIÓ DEL PROGRAMA D'ATENCIÓ DENTAL DEL PAÍS BASC

El Programa d'Atenció Dental Infantil (PADI) del País Basc, que porta en funcionament des de l'any 1990, constitueix una experiència de referència a Espanya ja que ofereix un nivell de cobertura pública pel que fa a l'atenció dental molt superior a la que s'observa a la resta de l'Estat. Aquest programa, a més de cobrir les extraccions com la resta de CA, inclou també una revisió anual i el tractament de càries i malformacions a tots els nens del País Basc d'entre 7 i 15 anys.

García (2005) va portar a terme una avaluació del PADI que pretenia escatir els efectes d'aquest programa sobre els tres *outcomes* següents: la probabilitat de no haver anat mai al dentista, de haver-hi anat en els darrers tres mesos i, finalment, que la darrera visita fos una revisió. L'estudi estima l'impacte del programa sobre aquestes variables utilitzant un model de dobles diferències. En concret, partint de dues edicions de l'Encuesta Nacional de Salud corresponents als anys 1987 i 2001, l'autora obté informació d'abans i de després de la introducció de la política tant per al grup de tractament (els nens del País Basc) com per al grup de comparació (els nens de la resta de CCAA). Els resultats obtinguts suggereixen que el programa només ha millorat un dels tres *outcomes* considerats: la probabilitat d'haver visitat el dentista en els darrers tres mesos, superior en el grup de tractament (nens del País Basc) respecte del de control (nens de la resta de CA).

Font: Elaboració pròpia a partir de García (2005).

3.6. ELECCIÓ ENTRE MÈTODES

Els apartats anteriors han posat de manifest l'existència de diversos mètodes susceptibles d'ésser utilitzats a l'hora de mirar d'establir l'impacte d'una determinada política. En general, una visió força compartida entre els avaluadors és que no existeix el mètode ideal, és a dir, un tipus de disseny en particular que, independentment de les circumstàncies, s'hauria d'aplicar de forma universal en totes les avaluacions d'impacte (Rossi, Lipsey i Freeman, 2004). A la pràctica, per tant, els avaluadors es veuen obligats a decidir entre diverses alternatives. Un element obvi que condiciona aquestes eleccions és la disponibilitat de temps i de recursos, però n'hi ha d'altres: les característiques del programa, la importància i l'ús que s'esperi fer dels resultats, la disponibilitat de dades, etc. Els apartats que segueixen tracten breument aquests aspectes, i argumenten a favor de la necessitat d'aproximar-se a l'elecció del mètode amb una mentalitat oberta, eclèctica i desproveïda d'apriorismes excessius.

3.6.1. CARACTERÍSTIQUES DEL PROGRAMA I DISPONIBILITAT DE DADES

Hi ha determinades característiques de les polítiques públiques que augmenten les possibilitats de mesurar amb rigor el seu impacte. Una especialment important és la relativa a la seva novetat i, més concretament, a la seva concepció com a prova pilot. En aquests casos, si es reuneixen una sèrie de condicions, com ara que la demanda potencial sigui superior a l'oferta o existeixin dubtes sobre l'efectivitat del programa, els experiments socials que utilitzen procediments d'assignació aleatoris poden constituir una forma d'avaluació

d'impacte a considerar. En qualsevol cas, malgrat que l'assignació no es produeixi de manera aleatòria, un programa pilot que s'implanti només en determinades zones geogràfiques obre les portes a dissenys no experimentals (*matching* o models DD) que utilitzin les àrees no-pilot per construir grups de comparació.

En qualsevol cas, fins i tot en aquells casos en què una nova política s'implementi sense proves pilots, afectant de sobte a tot el territori, segueix havent-t'hi possibilitats de construir grups de comparació si, pels motius que sigui, no tota la població potencialment beneficiària acaba participant en el programa. El pitjor dels casos es produeix, des de la perspectiva de l'avaluació d'impacte, quan una nova política s'implanta a escala nacional i afecta tota la població, ja que això només permet l'aplicació de mètodes reflexius (abans-després i sèries temporals).

Un altre avantatge de les polítiques noves, es materialitzin o no mitjançant proves pilot, és que permeten la introducció d'elements d'avaluabilitat mentre es desenvolupa la fase de disseny del programa. Com hem mencionat anteriorment, una avaluació d'impacte és, per definició, una avaluació *ex-post*, però les millors avaluacions d'impacte són aquelles que es planifiquen *ex-ante*. La possibilitat més extrema és que el mateix desplegament de la política es realitzi pensant en l'avaluació, com és el cas d'un experiment social, però de vegades n'hi ha prou de planificar una bona recollida de dades abans i després de la intervenció, que afecti sengles mostres de potencials beneficiaris i no beneficiaris, per incrementar enormement les possibilitats d'obtenir estimacions d'impacte creïbles mitjançant mètodes no experimentals.

Sovint però, l'impacte que es desitja avaluar no és el d'una política nova. En aquests casos, com que resulta impossible influir en "clau avaluadora" sobre el disseny del programa, el repte de l'avaluació consisteix a trobar característiques de la política i fonts d'informació que facin possible l'aplicació de les tècniques quasiexperimentals descrites en aquesta guia.

Així doncs, pel que fa a les característiques del programa, cal buscar-hi elements que possibilitin la construcció de contrafactuals: per exemple, si pels motius que sigui un determinat programa té llistes d'espera, els individus que en formen part poden constituir un grup de control natural respecte del qual estimar l'impacte del programa; així mateix, en la mesura en què existeixi variabilitat geogràfica en el grau d'implantació d'una política, les unitats territorials que disposin del programa poden comparar-se amb les que no en disposin (les comunitats autònomes poden constituir, en el cas d'algunes polítiques, una font de variabilitat a explorar en aquest sentit).

D'altra banda, respecte de la disponibilitat de fonts d'informació, constitueix una impressió general en el nostre país la infrautilització dels registres administratius amb finalitats avaluadores. En aquest sentit, un cop es té clar el disseny que pot prendre l'avaluació de la política o programa, la tasca de l'equip avaluador consisteix a identificar totes aquelles bases de dades amb informació rellevant sobre els individus que componen els grups de control i tractament prèviament definits, idealment amb l'horitzó temporal més ampli possible. Igualment, a més dels registres administratius, la recerca d'informació pot estendre's a enquestes ja disponibles o, fins i tot, a l'elaboració d'alguna de nova.

3.6.2. ECLECTICISME

Hi ha força casos en què l'equip avaluador, un cop explorades les característiques del programa i les fonts de dades disponibles, s'adonarà que poden utilitzar-se diverses de les tècniques quasiexperimentals comentades en els apartats previs i no una de sola. En aquestes quasiexperimentals, tret dels dissenys que no utilitzen grups de comparació, poc recomanables com ja s'ha comentat, no existeix evidència concloent que hi hagi una determinada metodologia que domini clarament la resta[9]. És per això que, en general, els avaluadors acostumen a aplicar simultàniament diversos tipus de metodologies, solució que permet addicionalment verificar fins a quin punt els resultats obtinguts depenen molt o poc de les eleccions de caràcter metodològic.

Les diverses tècniques en què hem centrat la nostra atenció fins el moment són metodologies d'anàlisi quantitatives. No és estranya la preeminència d'aquest tipus d'enfocament en l'avaluació d'impacte, ja que la qüestió fonamental a resoldre, que no és altra que la construcció d'un contrafactual, és de naturalesa bàsicament quantitativa. No obstant això, existeix una percepció creixent per part dels avaluadors que, per tal de millorar la robustesa de l'avaluació d'impacte, resulta recomanable complementar l'anàlisi utilitzant tècniques qualitatives (p. ex., entrevistes en profunditat o grups de discussió). El valor afegit que pot aportar llur utilització és permetre a l'equip avaluador millorar el seu coneixement sobre les condicions en què realment opera el programa, les perspectives dels seus beneficiaris, i d'altres elements fonamentals a l'hora d'entendre realment el perquè de l'impacte d'una política o programa (o de la seva absència).

Notes:

1 *El lector interessat pot aprofundir en l'estudi d'aquests mètodes seguint les lectures recomanades que apareixen en l'annex d'aquesta guia. També hi trobarà referències que tracten sobre dues tècniques que, donat el seu caràcter més tècnic, hem optat per deixar fora d'una guia de caràcter introductori: el models amb variables instrumentals i el disseny de regressió discontínua.*

2 *A llarg de l'exposició, ens referirem de manera genèrica a individus tractats i controls, tot i que hi ha força situacions en què la unitat d'anàlisi no són persones. És el que passaria, per exemple, si volguéssim avaluar una política d'incentius fiscals destinats a empreses per tal d'augmentar la seva recerca en R+D+I, o una reforma que donés més autonomia de gestió als centres escolars.*

3 *No entrarem en els detalls relatius a la grandària (nombre de persones) que han de tenir les mostres que componen els grups de control i tractament, ja que es tracta d'una qüestió força tècnica. Només mencionarem que quant més gran sigui la mida d'aquestes mostres, més possibilitats hi ha de detectar l'existència d'efectes atribuïbles a la política per petits que siguin. Vegeu per a una discussió detallada d'aquestes qüestions Purdon (2002).*

4 *Un llistat molt ampli d'avaluacions d'impacte fetes arreu, tant amb dissenys experimentals com quasi-experimentals, pot trobar-se a la web del Banc Mundial que apareix referenciada a l'annex d'aquesta guia.*

5 *L'elevat cost d'un experiment no constitueix, per si sol, un argument suficient per decidir no portar-lo a terme. La comparació rellevant s'ha de realitzar tenint en compte també les conseqüències que pot suposar estendre una política que, malgrat no tenir cap impacte demostrat, absorbeix una quantitat ingent de recursos públics.*

6 *Els models d'elecció discreta són aquells que pretenen establir la relació existent entre una variable dependent binària (p. ex. participar o no) i un seguit de variables independents que a priori es considera poden influir sobre aquella. La diferència entre els dos models esmentats rau en la forma funcional que es suposa relaciona la variable dependent amb les independents: una funció logística en el cas del lògit, una funció normal en el cas del pròbit. Vegeu Corbetta (2007) per a més detalls sobre aquest tipus de models.*

7 *Aquests tipus de models es coneixen en anglès amb el nom de difference-in-differences, tot i que sovint s'utilitza l'abreviatura diff-in-diff per referir-s'hi. Hem optat per traduir-los per "models de dobles diferències" seguint la proposta de traducció al castellà suggerida per Vera-Hernández (2003).*

8 *És important assenyalar que, a l'hora d'estimar impactes mitjançant un model DD, no cal que la informació sigui longitudinal (això és, sobre els mateixos individus abans i després de la intervenció). Poden fer-se servir dades de secció creuada (dues enquestes realitzades abans i després de la intervenció a individus diferents), sempre i quan puguem identificar beneficiaris i no beneficiaris en un i altre moment.*

9 *La manera com s'avalua la robustesa dels mètodes d'avaluació d'impacte quasiexperimentals és aplicant-los a bases de dades que han estat obtingudes a partir d'un experiment social. Així doncs, partint de la premissa que l'experiment social permet identificar l'impacte real, els resultats obtinguts per la resta de mètodes es comparen amb aquests.*

BIBLIOGRAFIA

BÉLAND, F. [et al.]. "A system of integrated care for older persons with disabilities in Canada: Results from a randomized controlled trial". *The Journals of Gerontology: Medical Sciences (2006)*, núm. 61 (4), p. 367-373.

CORBETTA, P. *Metodología y Técnicas de Investigación Social*. Madrid: MacGrawHill, 2007.

GARCÍA, P. "Evaluación de un Programa de Atención Dental Público: PADI en el País Vasco". *Ekonomiaz (2005)*, núm. 60, p. 62-89.

HECKMAN, J.; HIDEHIKO, I.; TODD, P. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies (1997)*, núm. 64 (4), p. 605-654.

KUEGER, A.B. *"Experimental Estimates of Education Production Functions". The Quarterly Journal of Economics (1999)*, núm. 114, p. 497-532.

MOFFITT, R. A. "The Role of Randomized Field Trials in Social Science Research. A Perspective from Evaluations of Reforms of Social Welfare Programs". *American Behavioral Scientist (2004)*, núm. 47 (5), p. 506–40.

PURDON, S. *Estimating the impact of labour market programmes*. Londres: Department for Work and Pensions, 2002. (Working Paper núm. 3)

RAVALLION, M. *Evaluating Anti-Poverty Programs*. Washington DC: World Bank, 2006. (Policy Research Working Paper 3625)

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

TOHARIA, L. [et al.]. *Estudio de evaluación de la formación ocupacional en Catalunya*. Barcelona: Servei d'Ocupació de Catalunya, 2008. (mimeo)

ANNEX. GUIA DE RECURSOS

MANUALS

MANUALS ESPECÍFICS D'AVALUACIÓ D'IMPACTE:

BAKER, J. *Evaluating the Impact of Development Projects on Poverty—A Handbook for Practitioners*. Washington, DC: World Bank, 2000.

ASIAN DEVELOPMENT BANK. *Impact Evaluation: Methodological and Operational Issues*. / Manila: ADB, 2006.
(http://www.adb.org/Documents/Handbooks/Impact-Analysis/default.asp)

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company, 2002.

MANUALS GENERALS D'AVALUACIÓ AMB CAPÍTOLS SOBRE AVALUACIÓ D'IMPACTE:

ROSSI, P. H.; LIPSEY, M. W; FREEMAN, H. E. *Evaluation: a systematic approach*. 7a ed. / Londres: Sage, 2004.

WEISS, C. *Evaluation.* 2a ed. New Jersey: Prentice Hall, 1998.

ARTICLES

La majoria d'articles que es mencionen a continuació, i d'altres de relacionats, poden descarregar-se gratuïtament des de la següent pàgina web del Banc Mundial:

http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20215333~menuPK:451260~pagePK:148956~piPK:216618~theSitePK:384329,00.html

ARTICLES INTRODUCTORIS:

RAVALLION, M. "The Mystery of the Vanishing Benefits. Ms Speedy Analyst´s Introduction to Evaluation". *World Bank Economic Review (2001)*, núm. 15, p. 115-140.

VERA-HERNÁNDEZ, M. "Evaluar intervenciones sanitarias sin experimentos". *Gaceta Sanitaria (2003)*, núm. 17, p. 238-248.
(http://scielo.isciii.es/pdf/gs/v17n3/revision.pdf)

ARTICLES QUE REVISEN DIVERSES TÈCNIQUES D'AVALUACIÓ:

BLUNDELL, R.; COSTA DIAS, M. "Evaluation methods for non-experimental data". *Fiscal Studies (2000)*, núm. 21(4), p. 427-468.

RAVALLION, M. *Evaluating Anti-Poverty Programs*. Washington DC: World Bank, 2006. (Policy Research Working Paper 3625)

ARTICLES SOBRE EXPERIMENTS SOCIALS:

BURTLESS, G. "The case for randomized field trials in economic and policy research". *Journal of Economic Perspectives (1995*), núm. 9, p. 63-84.

DUFLO, E.; GLENNERSTER, R.; KREMER, M. U*sing Randomization in Development Economics Research: A Toolkit*. Londres: CEPR, 2007. (CEPR working paper, number 6059)

ARTICLES SOBRE *MATCHING*:

CALIENDO, M.; KOPEINIG, S. "Some Practical Guidance for the Implementation of Propensity Score Matching". *Journal of Economic Surveys (2008)*, núm. 22, p. 31-72.

IMBENS, G. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". *The Review of Economic and Statistics (2004)*, núm. 86, p. 4-29.

ARTICLES SOBRE VARIABLES INSTRUMENTALS:

HECKMAN, H. "Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations". *Journal of Human Resources (1997)*, núm. 32, p. 441-462.

ARTICLES SOBRE REGRESSIÓ DISCONTÍNUA:

LEE, D.; LEMIEUX, T. *Regression Discontinuity Designs in Economics*. Boston: NBER, 2009. (Working Paper Series, núm. 14723)

ENLLAÇOS D'INTERÈS

Network of Networks on Impact Evaluation (NONIE) http://www.worldbank.org/ieg/nonie/index.html

Banc de Desenvolupament Iberoamericà
http://www.iadb.org/ove/DefaultNoCache.aspx?Action=WUCPublications@ImpactEvaluations

Avaluacions d'impacte a Colòmbia
http://www.dnp.gov.co/PortalWeb/Programas/Sinergia/EvaluacionesEstrat%C3%A9gicas/tabid/215/Default.aspx

Avaluacions d'impacte a Xile
http://www.dipres.cl/572/propertyvalue-15223.html

Base de dades del Banc Mundial sobre avaluacions d'impacte
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21534261~menuPK:412159~pagePK:210058~piPK:210062~theSitePK:384329,00.html

ivàlua
Institut Català d'Avaluació de Polítiques Públiques
Gran Via de les Corts Catalanes, 680, 3r
08010 Barcelona
T. +34 93 554 53 00
info@ivalua.cat
www.ivalua.cat

Exemplar imprès: 12€
Versió electrònica: www.ivalua.cat