

Metodo sperimentale, approccio controfattuale e valutazione degli effetti delle politiche pubbliche

Alberto Martini

Nell'articolo "Dieci anni di RIV", pubblicato sull'ultimo numero di questa rivista, Claudio Bezzi sollecita una discussione sulla rilevanza dell'approccio (o metodo) sperimentale, sottolineando la contraddizione tra l'importanza che la teoria assegna al metodo sperimentale e l'apparente mancanza di esperienze italiane di valutazione sperimentale (e quasi-sperimentale):

"L'approccio sperimentale: è una sorta di *must* della ricerca valutativa. Se siamo una rivista aperta, come asseriamo, non possiamo dividerci fra sostenitori e osteggiatori di questo approccio; esso rappresenta un pilastro di una fetta rilevante della teoria valutativa contemporanea, e alcuni dei maggiori autori la ritengono tuttora la regina degli approcci. E comunque, anche il negarne la validità con altrettanta competenza teoretica, non segnala altro che l'importanza del tema. Mi scuseranno gli autori a me sconosciuti, faccio ammenda, ma anche qui mi viene da chiedere: dove sono le esperienze *italiane* di valutazione sperimentale o quasi sperimentale? Perché qualcuno non scrive a favore (o contro) tale approccio sostenendo anche empiricamente le proprie tesi?"

Raccogliendo questo invito, tenterò di argomentare le seguenti tre tesi.

- (a) L'importanza del metodo sperimentale non va esagerata: per collocarlo nella giusta prospettiva, è opportuno considerare questo metodo come *caso speciale* di un approccio più generale alla valutazione, l'*approccio controfattuale*, che si serve di metodi sperimentali ma anche e soprattutto di metodi non-sperimentali.
- (b) A sua volta, l'approccio controfattuale non è rilevante per qualsiasi problema di valutazione, bensì solo per un tipo particolare di domanda valutativa, quella che chiede di quantificare l'effetto di un particolare intervento pubblico. L'approccio controfattuale, e *a fortiori* il metodo sperimentale, possono essere pacificamente ignorati da chi non debba rispondere a questo tipo di domanda valutativa. Viceversa, chi sostiene di fare valutazione degli effetti (o dell'impatto) dovrebbe confrontarsi *seriamente* con questo approccio e capirne sia i limiti sia le potenzialità, per poi eventualmente "negarne la validità con altrettanta competenza teoretica": ma non limitarsi a liquidarlo come irrilevante o peggio a ignorarlo, preferendovi l'uso (metodologicamente indifendibile) di qualche "indicatore di impatto".
- (c) La ragione della indubbia scarsità di esperienze di valutazioni di ispirazione "controfattuale" in Italia è da ricercarsi non tanto in obiezioni di tipo etico o teoretico (rilevanti solo in alcuni casi) o in difficoltà di ordine tecnico o logistico (superabili con adeguate risorse), bensì nella scarsa propensione delle pubbliche amministrazioni italiane a mettere *genuinamente* in discussione l'efficacia dei propri interventi e a volerne quindi *genuinamente* quantificare gli effetti. Atteggiamento in aperto contrasto con quello tenuto negli ultimi decenni dalle loro controparti nord-americane, alle cui concrete e genuine richieste di valutazione degli effetti delle politiche è legato l'ampio utilizzo del metodo sperimentale (nonché il continuo sviluppo di metodi non-sperimentali rigorosi e robusti.)

1. Davvero irrilevante?

L'invito di Claudio Bezzi contiene un ossequio formale all'importanza assegnata dalla teoria al metodo sperimentale: “un *must* della ricerca valutativa”, “un pilastro di una fetta rilevante della teoria valutativa contemporanea”. Questo tipo di ossequio formale al metodo sperimentale si incontra spesso nella letteratura italiana ed europea, accompagnato però da forti dubbi sulla sua *rilevanza pratica*. Un esempio di questo dubbio ce lo offre lo stesso Claudio Bezzi, che nel suo manuale scrive:

“Benché il disegno sperimentale sia contemplato in tutta la letteratura metodologica, non ci si può nascondere la sua difficile realizzabilità *di fatto* quando si lavora nell'ambito delle scienze sociali. La ricerca sperimentale (...) presuppone infatti il controllo ferreo della clausola *ceteris paribus*, realizzabile sostanzialmente (chiudendo almeno un occhio) in piccoli gruppi e con esperimenti di laboratorio, di cui è lecito dubitare la riproducibilità in un contesto *reale*. (...) I disegni sperimentali sono essenzialmente dei modelli di studio, finanziati e realizzati probabilmente in ambiente universitario, e non dei modelli operativi finanziati e realizzati da committenti impegnati in un programma *reale*” (...) Una ricerca sperimentale pura sembra, in estrema sintesi, assai difficilmente realizzabile.¹

Questa breve citazione non fa giustizia delle complesse argomentazioni che Bezzi presenta nel suo manuale, ma è sufficiente come esempio della convinzione diffusa che il metodo sperimentale sia irrilevante per la pratica valutativa e sia destinato a restare tale, non per scelta, ma per colpa del metodo stesso.

Per chi conosce a fondo l'esperienza statunitense, questa convinzione risulta a dir poco sorprendente. Scrive Larry Orr, uno dei più affermati e competenti *practitioner* di questo tipo di valutazione negli Stati Uniti, in una sua monografia non recentissima ma ancora attuale:

“Nei trent'anni in cui mi sono occupato di esperimenti sociali, questo approccio è divenuto sempre più il *gold standard* della *program evaluation* (...) Questo insieme consolidato di metodi è scaturito dal grande numero di esperimenti realizzati durante questo periodo (...) Su questi metodi molto poco è stato pubblicato nelle riviste accademiche.”²

Le affermazioni di Bezzi e di Orr sono in evidente contrasto: sembra che i due autori parlino di cose diverse. In realtà parlano della stessa cosa, ma da due prospettive molto diverse. Diverso è soprattutto ciò che i due autori intendono per “valutazione”. Orr si rivolge ad un *audience* che di mestiere si occupa di una particolare tipologia di *program evaluation*, intesa come “quantificazione degli effetti prodotti da un ben definito intervento pubblico su un ben definito gruppo di beneficiari”. Esempi di questo tipo di valutazione di interventi sociali, di cui Orr e la sua *audience* si occupano, sono:

- valutare (quantificare) l'effetto, sul reddito da lavoro e sul grado di autosufficienza economica, di corsi di riqualificazione professionale destinati ai percettori di sussidi di *welfare*;
- valutare (quantificare) l'effetto, sulla durata della disoccupazione, dell'assistenza alla ricerca di lavoro fornita a lavoratori licenziati in seguito a ristrutturazioni;
- valutare (quantificare) l'effetto, sul loro tasso di occupazione futuro, dell'erogazione di borse-lavoro a favore di ex-detenuti;
- valutare (quantificare) l'effetto, sulla performance degli alunni nei test standardizzati, della riduzione del numero di alunni per classe;
- valutare (quantificare) l'effetto, sul consumo di farmaci, del ticket sui medicinali.

¹ Claudio Bezzi, *Il disegno della ricerca valutativa*, seconda edizione, Milano: Franco Angeli, 2003, pagg. 336-340. Le altre tesi che su questo tema Bezzi sviluppa nel suo manuale riguardano sostanzialmente la *limitata applicabilità* del metodo sperimentale: tesi con cui siamo pienamente d'accordo, come argenteremo più avanti.

² Larry Orr, *Social Experiments: Evaluating Public Programs with Experimental Methods*, London: Sage, 1999, pag. 2.

La nostra enfasi sul “quantificare” come specificazione del “valutare” *non* intende assolutamente affermare la superiorità dei metodi quantitativi sui metodi qualitativi, una contrapposizione che riteniamo priva di ragion d’essere. Al contrario, la nostra enfasi intende sottolineare la *ristrettezza* della prospettiva in cui si colloca il tipo di ricerca valutativa appena esemplificato, a sua volta motivata da un *particolare* tipo di domanda da parte della committenza pubblica, domanda che può essere stilizzata così: “quell’intervento (*program*), messo in campo per alleviare quella ben precisa problematica sociale, ottiene (nel senso di “causa”) i risultati sperati e quindi *funziona?* (*does it work?*)”

Si può dissentire, per ragioni ideologiche e/o metodologiche, dall’opportunità di porsi questo tipo di domanda valutativa, ma un fatto resta comunque: l’utilizzo del metodo sperimentale negli Stati Uniti è motivato dal persistere e dal rinnovarsi di questo tipo di domanda, da parte della committenza pubblica, riguardo al successo di particolari politiche e interventi nel campo educativo, sanitario, del lavoro e soprattutto del *welfare* (inteso come assistenza economica alle famiglie povere). Termini quali “*randomized controlled trial*” o “*random assignment evaluation*”, o il meno tecnico “*social experiment*”, sono entrati nel linguaggio delle amministrazioni pubbliche americane che, a livello federale o di singolo stato, disegnano e gestiscono questo tipo di politiche. Si dà il caso persino di alcune leggi che danno mandato all’Esecutivo di valutare determinati interventi mediante esperimenti sociali.³

Tale pratica valutativa ha visto un coinvolgimento solo marginale delle istituzioni accademiche, innanzitutto perché la conduzione di un esperimento sociale richiede uno sforzo logistico e organizzativo che mal si adatta alle piccole dimensioni di un tipico dipartimento o centro di ricerca universitario.

Questa pratica valutativa la si ritrova però ampiamente codificata nei trattati teorici di valutazione, scritti prevalentemente da accademici, dalla cui lettura Bezzi e altri ricavano l’impressione che il metodo sperimentale sia “un *must* della ricerca valutativa”. Ma se questo è vero, non è tanto perché questo metodo rappresenti *in astratto* “un pilastro di una fetta rilevante della teoria valutativa contemporanea”; bensì più semplicemente perché, nel Paese in cui viene prodotta una parte preponderante della teoria valutativa contemporanea, una quota significativa delle domande poste dai “committenti impegnati in un programma reale” riguardano appunto la quantificazione dei suoi effetti.⁴

2. La “regina degli approcci”: ma di quali approcci, esattamente?

Un altro errore in cui incorrono spesso i critici del metodo sperimentale è quello di non rendersi conto che questo metodo rappresenta un caso particolare di un più ampio approccio, il *paradigma controfattuale*: arrivando talvolta ad ignorare *in toto* l’esistenza dell’ampio patrimonio di contributi metodologici motivati appunto dalla consapevolezza che, essendo il puro metodo sperimentale applicabile in situazioni limitate, per quantificare gli effetti di una politica occorra spesso ricorrere a metodi non-sperimentali.

Sia i metodi sperimentali sia quelli non-sperimentali partono da una comune *definizione di effetto*: “effetto di un intervento è la differenza tra ciò che osserviamo *dopo* che l’intervento è stato attuato e ciò che avremmo osservato, nello stesso periodo e per gli stessi soggetti, *in assenza* di

³ Per una rassegna esaustiva degli esperimenti condotti fino alla metà degli anni '90, un’utile fonte è David Greenberg e Mark Shroder, *Digest of Social Experiments*, Washington: The Urban Institute Press, 1997.

⁴ A questa domanda valutativa da parte delle amministrazioni pubbliche statunitensi non fa sempre seguito un utilizzo pieno e diretto dei risultati delle valutazioni. Ma questa è un’altra questione, molto più complessa. Resta il fatto che le amministrazioni pubbliche statunitensi (e alcune fondazioni filantropiche) hanno investito, a partire dalla seconda metà degli anni sessanta, quantità notevoli di risorse nella realizzazione di esperimenti sociali. Per una discussione su questo tema si veda il saggio di David Greenberg, Donna Links e Marvin Mandell; *Social Experimentation and Public Policymaking*, Washington: The Urban Institute Press, 2003.

intervento”. Quindi effetto è *definibile* come differenza (rispetto alle variabili su cui la politica pubblica intende incidere) tra un valore *osservabile* e uno *ipotetico*, per sua natura *non osservabile*.

La non-osservabilità del controfattuale venne definita, ormai vent’anni fa, dallo statistico Paul Holland come “*the fundamental problem of causal inference*”.⁵ Il riferimento alla *causal inference* mette bene in evidenza l’obiettivo conoscitivo dell’approccio controfattuale: stabilire l’esistenza di un *legame causale* tra la realizzazione di un intervento e ciò che si osserva tra i destinatari di quell’intervento. Identificare cioè il contributo netto dell’intervento, separandolo dai molteplici fattori, estranei all’intervento, che influenzano comunque i destinatari e le loro condizioni o comportamenti. Tale separazione ha lo scopo di capire se i cambiamenti che si osservano tra i destinatari sono “merito” dell’intervento⁶ (e le risorse ad esso dedicate sono “ben spese”) o non sono dovuti piuttosto a “tutto il resto che cambia” (l’intervento quindi non ha meriti da vantare e i soldi spesi sono *deadweight*). Non è questo l’unico problema valutativo, e forse neanche il più importante: è però un problema che sta a cuore (almeno a parole) ad alcuni committenti pubblici e quindi andrebbe affrontato con il dovuto rigore e onestà intellettuale nel dare una risposta alle loro domande.

Dalla non-osservabilità del controfattuale discende come conseguenza logica la non-osservabilità dell’effetto. A rigore, un effetto *non può mai essere osservato* (né quindi “misurato”) direttamente, perché non è possibile osservare *contemporaneamente* gli *stessi* soggetti nello *status* di beneficiari di un intervento e in quello di non-beneficiari.⁷

Veniamo al punto cruciale: il fatto che un effetto non sia mai osservabile direttamente non elimina però la possibilità di *argomentare* qualcosa di plausibile circa tale effetto. Nella misura in cui la situazione controfattuale possa essere *plausibilmente* ricostruita con *altre* informazioni, si può comunque stimare l’effetto come differenza tra la situazione osservata post-intervento e la (plausibile) ricostruzione della situazione controfattuale. La partita qui si gioca in termini di maggiore o minore *plausibilità*, non di “verità scientifica”.

Metodo sperimentale come caso speciale

Qui, e fondamentalmente qui, entra in scena il metodo sperimentale, come caso speciale dell’approccio controfattuale: speciale perché con questo metodo la situazione controfattuale viene ricostruita *osservando ciò che succede ad un gruppo di controllo* composto da soggetti *molto simili* a quelli esposti all’intervento (questi ultimi sono detti collettivamente *gruppo sperimentale*). Tale ricostruzione viene ritenuta particolarmente *plausibile* per l’assenza di differenze di partenza tra i due gruppi: in ciò sta l’essenza della “superiorità” del metodo sperimentale.

La similitudine tra i due gruppi è conseguenza del fatto che sono scelti mediante assegnazione casuale (o randomizzazione), cioè in pratica mediante una qualche forma di sorteggio: il che a sua volta presuppone che il valutatore sia in grado di *manipolare* il processo attraverso cui i destinatari potenziali di un intervento pubblico accedono effettivamente alle prestazioni in cui questo intervento consiste.

La necessità di manipolare il processo di selezione mediante assegnazione casuale è al tempo stesso la grande forza e il grande limite del metodo sperimentale. *Forza* perché l’assegnazione casuale garantisce che (a parte alcune complicazioni) l’unica differenza tra gruppo sperimentale e gruppo di controllo stia nel fatto di essere o meno esposti all’intervento e che quindi

⁵ Paul W. Holland, “Statistics and Causal Inference”, *Journal of the American Statistical Association*, Vol. 81, No. 396, pp. 945-960, 1986.

⁶ O “colpa” dell’intervento, nel caso di effetti “non desiderati”. La desiderabilità o meno degli effetti non ha di per sé nessuna conseguenza dal punto di vista metodologico, si tratta in entrambi i casi di stabilire un legame causale. Può avere conseguenze sulla disponibilità di dati, in quanto un effetto non desiderato è spesso anche un effetto *non atteso*.

⁷ Questa semplice argomentazione è sistematicamente ignorata da coloro che illudono sé stessi (e gli altri) sostenendo di poter “misurare” gli effetti di una politica mediante “indicatori di impatto”. Come un semplice indicatore possa risolvere “*the fundamental problem of causal inference*” resta, per chi scrive, un mistero.

tutto ciò che succede al gruppo di controllo riproduce *plausibilmente* ciò che sarebbe successo al gruppo sperimentale *se questo non fosse stato esposto* all'intervento.

La necessità di manipolare il processo di selezione rappresenta anche il grande *limite* di questo metodo, fondamentalmente per due ragioni:

- (i) perché in molti casi le caratteristiche stesse dell'intervento *non consentono* tale manipolazione, pena alterarne in modo fondamentale il funzionamento. Si pensi al caso dell'imposizione di un obbligo di legge, quale quello di indossare il casco alla guida di un ciclomotore: sarebbe legalmente oltre che materialmente impossibile applicare selettivamente questo obbligo ad alcuni e non ad altri guidatori;
- (ii) l'assegnazione casuale implica il più delle volte negare una prestazione ad alcuni soggetti che hanno il diritto o anche solo l'aspettativa di riceverla: tale esclusione non solo solleva obiezioni di tipo etico (che gli esperimenti sociali condividono con la sperimentazione clinica), ma soprattutto crea *grosse difficoltà nell'ottenere il consenso* dei vari attori della politica pubblica, che possono non condividere le motivazioni alla base della domanda di valutazione. Si pensi al caso in cui, per verificare sperimentalmente l'efficacia di un intervento di recupero effettuato dai SerT di una certa regione, si intendesse sottoporre al trattamento solo una parte, per giunta selezionata a caso, dei tossicodipendenti che si rivolgono al servizio: si pensi alle difficoltà che si incontrerebbero nell'ottenere il consenso di tutti gli attori coinvolti, dagli operatori dei SerT all'Assessore regionale alla Sanità!

Quando tale manipolazione sia impossibile o comunque non sia proponibile, ma si desideri comunque quantificare l'effetto di un intervento, occorre ricorrere ad altri metodi per ricostruire la situazione controfattuale, metodi che i valutatori di formazione statistico-economica definiscono come *non-sperimentali*, mentre quelli di formazione sociologica o psicologica definiscono *quasi-sperimentali*, in omaggio alla terminologia introdotta quarant'anni fa da Donald Campbell.⁸ Quale sia l'etichetta preferita, qui ci troviamo di fronte a una grande varietà di metodi, la cui applicabilità dipende dalle caratteristiche dell'intervento, dai dati utilizzabili, dalle risorse a disposizione e dal tempo entro cui si deve fornire una risposta al committente.

Argomentare la plausibilità

Non è questa la sede adatta per una rassegna dei metodi non-sperimentali oggi utilizzati nella valutazione degli effetti delle politiche, per cui si rimanda ad altre fonti.⁹ Ciò che qui è opportuno sottolineare è che tutti questi metodi richiedono, per ottenere la stima dell'effetto, di imporre assunti non testabili empiricamente: la non-testabilità di questi assunti impone al valutatore l'onere di argomentarne la plausibilità caso per caso, in quanto dalla loro plausibilità dipende quella della stima dell'effetto ottenuta.¹⁰

⁸ Autore di uno dei primi e fondamentali lavori sul tema: Donald Campbell e Julian Stanley, "*Experimental and Quasi-experimental Designs for Research*", Chicago: Rand McNally, 1966.

⁹ Per una rassegna esaustiva ma molto tecnica si veda: James Heckman, Robert LaLonde e Jeffrey Smith, "The Economics and Econometrics of Active Labor Market Programs", in Orley Ashenfelter e David Card (a cura di), *Handbook of Labor Economics*, pagg. 1865-2097, North-Holland, 1999. Per una rassegna meno tecnica e scritta da europei e pubblicata in Europa, si veda Richard Blundell e Monica Costa Dias, "Alternative approaches to evaluation in empirical microeconomics", *Portuguese Economic Journal*, Vol. 1, No. 2, pagg. 91-115, 2002. Per una rassegna non esaustiva ma agli antipodi rispetto alle precedenti in termini di tecnicismo si veda Alberto Martini, Luca Mo Costabella e Marco Sisti, "Valutare gli effetti delle politiche pubbliche: metodi e applicazioni al caso italiano", di prossima pubblicazione in *Materiali Formez*: oltre ad una esposizione semplificata della metodologia, questo lavoro contiene sei studi di caso di valutazioni non-sperimentali condotte in Italia negli ultimi anni.

¹⁰ Ad onor del vero, anche il metodo sperimentale richiede di imporre assunti non testabili, ad esempio quello che la randomizzazione non alteri *di per sé* il comportamento del gruppo sperimentale e/o di quello di controllo. Se gli assunti necessari per la validità del metodo sperimentale siano in generale più plausibili degli assunti necessari per la validità dei vari metodi non sperimentali, è ancora oggetto di acceso dibattito all'interno della comunità di valutatori e

L'esempio più elementare della necessità di argomentare la plausibilità degli assunti sta nell'utilizzo, a scopo di stima degli effetti, della semplice differenza pre-post. Incidentalmente, va notato come questa differenza rappresenti, nell'immaginario collettivo, la definizione stessa di effetto. Chiedete a una persona di livello culturale medio di definire "effetto di un intervento" e vi verrà data, nella migliore delle ipotesi, una definizione di questo tenore: "la differenza tra ciò che si osserva *dopo* che l'intervento è stato attuato e ciò che si osservava *prima* dell'intervento".

Secondo la logica controfattuale, questa non è la definizione di effetto, bensì l'applicazione del più semplice, immediato e al tempo stesso potenzialmente fallace, dei metodi non-sperimentali: la *differenza pre-post*. Differenza che può essere utilizzata, senza incorrere in obiezione alcuna, come semplice strumento descrittivo: tale era il livello della variabile di interesse (o indicatore) prima dell'intervento, tale è il livello dopo, quindi tale è l'aumento o la diminuzione. Senza alcuna implicazione causale.

Maggiori cautele sono richieste invece quando si voglia *interpretare* la differenza pre-post come effetto dell'intervento: questa interpretazione è plausibile esclusivamente nella misura in cui la situazione pre-intervento rappresenti un'approssimazione plausibile della situazione controfattuale. A sua volta, questa approssimazione dipende dall'assunto, non testabile empiricamente con sole due osservazioni, che in assenza di intervento la situazione dei beneficiari *non sarebbe cambiata*, e quindi il fenomeno su cui l'intervento vuole incidere abbia una *dinamica nulla*. Nelle situazioni specifiche in cui l'assunto di "dinamica nulla" non è ritenuto plausibile, non è neppure da ritenersi plausibile la stima dell'effetto derivata calcolando la differenza pre-post.

Concludendo su questo punto, se il metodo sperimentale è una regina, lo è solo dell'approccio controfattuale, e forse neppure di questo, poiché alcuni molto autorevoli sostenitori di questo approccio contestano in modo veemente l'assolutezza del suo regno. A sua volta, l'approccio controfattuale (che si serve ampiamente anche di metodi non-sperimentali) non ha alcuna ambizione al titolo di regina della valutazione, in quanto i suoi proponenti ne riconoscono onestamente non soli i limiti cognitivi e applicativi, ma soprattutto il fatto di rappresentare solamente *un modo* per rispondere ad *una specifica domanda* di valutazione, quella sulla quantificazione di effetti (a cui si voglia dare un'interpretazione causale).

La raccomandazione per il valutatore quindi è: "*Se non ti poni (o non ti pongono) questa domanda, dimenticati senza problemi esperimenti e controfattuale, e valuta felice e contento.*"

3. The GUIDE e l'insostenibile leggerezza degli indicatori

I sostenitori¹¹ dell'approccio controfattuale ambirebbero, invece che al titolo di "regina della valutazione", a qualcosa di molto più modesto: che il loro sia riconosciuto come *un* impianto rigoroso e coerente per impostare *un* particolare problema valutativo. Non il solo approccio possibile, ma uno che andrebbe preso *seriamente* in considerazione per affrontare quel particolare problema valutativo. Questo approccio viene invece totalmente ignorato o trattato con superficialità da una parte molto significativa di quei valutatori e committenti (italiani ed europei) che pure si pongono l'obiettivo ambizioso di quantificare gli effetti di politiche pubbliche. Tale scelta è quasi la regola nei manuali prodotti per la valutazione dei fondi strutturali europei, nei quali peraltro regna sovrano il richiamo all'importanza di quantificare effetti, impatti, esiti, risultati.

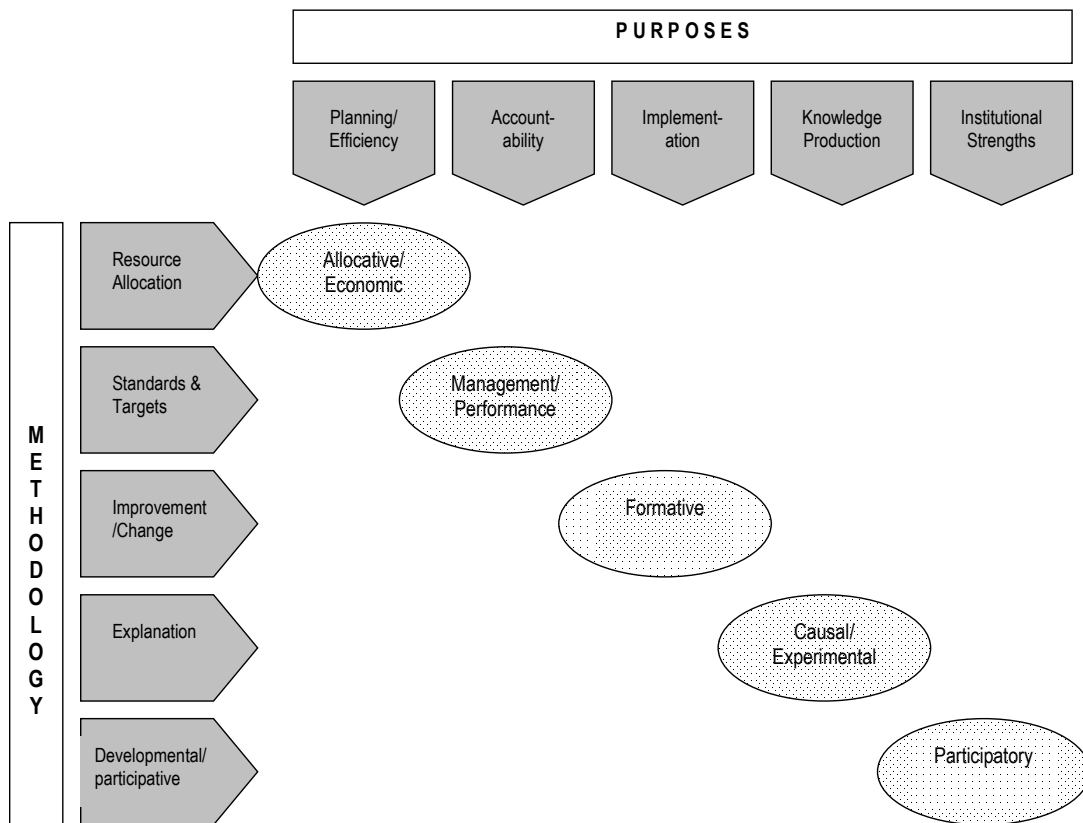
studiosi che si rifanno all'approccio controfattuale. Per un esempio di contributo a tale acceso dibattito, si veda James Heckman e Jeffrey Smith, "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, 1995.

¹¹ Molti e autorevoli sono i sostenitori di questo approccio, divisi su molte cose (ad esempio sulla priorità da dare al metodo sperimentale) ma concordi sulla necessità del rigore metodologico per valutare gli effetti delle politiche pubbliche. Citiamo qui solo tre tra i più noti: James Heckman, economista dell'Università di Chicago, premio Nobel per l'economia nel 2000; Donald Rubin, statistico dell'Università di Harvard, da cui prende il nome il modello di Rubin, una formalizzazione rigorosa dell'approccio controfattuale; Judith Gueron, per 20 anni presidente della Manpower Demonstration Research Corporation, il centro di ricerca *non-profit* che più ha contribuito alla realizzazione di esperimenti sociali negli Stati Uniti e in Canada.

Prendiamo come esempio emblematico il documento metodologico prodotto dall'Unione Europea, *"The Evaluation of Socio-Economic Development: The GUIDE"* pubblicato nel 2003.¹² Questo documento si occupa di valutazione a 360 gradi, identificando 5 diversi scopi della valutazione, 5 diverse posizioni metodologiche e 5 diverse tipologie prevalenti di valutazione. A noi interessa esclusivamente una piccola parte di questo universo, cioè la valutazione che ha come obiettivo esplicito *"identifying outputs, outcomes and impacts."* A questo proposito, *The GUIDE* è molto enfatica: *"For many policy makers, identifying, describing and quantifying such outputs, outcomes and impacts is a major benefit of evaluation."* (pag. 9) Prodotti, risultati, impatti, vanno tutti identificati, descritti e quantificati, e tutto ciò sarebbe, per i *policy-maker*, una funzione essenziale della valutazione.

Alla luce di queste affermazioni, ci aspetteremmo una discussione approfondita di come tali ambizioni risultati conoscitivi possano essere ottenuti. La discussione metodologica svolta nella *GUIDE* è senza dubbio molto elaborata. Il grafico seguente mostra come i 5 diversi scopi della valutazione si relazionino con le 5 diverse posizioni metodologiche per dare luogo alle 5 diverse tipologie prevalenti di valutazione.

Box 1.3 The two axes of evaluation, Purpose and Methodology



Riprodotta da *"The Evaluation of Socio Economic Development: The GUIDE"* (pag. 22)

All'intersezione del quarto scopo, indicato come *"knowledge production"*, e della quarta metodologia, indicata come *"explanation"*, troviamo, non senza un po' di sorpresa, il termine *causal/experimental*. *The GUIDE* sembra quindi assegnare al metodo sperimentale il ruolo di una

¹² Curato per conto dell'UE dal Tavistock Institute con la collaborazione di GHK Consulting Ltd e dell'Istituto per la Ricerca Sociale. Scaricabile per intero dal sito www.evaled.info.

delle 5 tipologie principali di valutazione, lusingandolo con una collocazione davvero invidiabile, all'incrocio tra "produzione di conoscenza" e "spiegazione".

Doppiamente delusi quindi restiamo nel constatare con quanta *superficialità* l'approccio *causal/experimental* viene trattato nel prosieguo del manuale. Ci saremmo aspettati una seria discussione dell'approccio controfattuale, dei limiti applicativi del metodo sperimentale, dei limiti conoscitivi dei metodi non-sperimentali. Questo non per semplice deferenza alla teoria valutativa contemporanea, ma per l'asserita importanza che avrebbe per i *policy-maker* quantificare gli effetti dei programmi socio-economici. Quello che troviamo invece è una sequenza di affermazioni superficiali, a nostro parere intrise di pregiudizi: ancora una volta, ad un iniziale ossequio formale fa seguito una stroncatura aprioristica e poco informata.

Ma cosa giustifica una critica così *tranchant* a un documento prodotto dall'Unione Europea? Le citazioni seguenti, non esaustive ma indicative, motivano questo atto di inconsueta irriverenza.

Primo esempio. Il passaggio seguente discute i metodi che sarebbero utilizzabili nel caso si volessero valutare interventi che hanno una finalità ben definita ("*more defined scope*"), cioè interventi mirati a particolari problematiche o sotto-popolazioni.

When an evaluation concerns interventions with a more defined scope, it is possible to carry out an in-depth analysis of the causal links between the intervention and its effects. Several techniques may be used in this context:

(...)

- *Comparison groups are used to estimate net effects by noting the difference between a group of beneficiaries and a group of non-beneficiaries.*
- *Regression analysis is used to estimate net effects and to determine whether the causal links between the intervention and its effects are statistically significant. (pag. 114)*

Quando abbiamo a che fare con interventi mirati, dice *The GUIDE*, è possibile analizzare in profondità il *legame causale tra l'intervento e i suoi effetti*. Gli accenni al "gruppi di confronto" e alla "regressione" rivelano però una nozione assai nebulosa dell'inferenza causale. Apprendiamo infatti che gli effetti netti¹³ possono essere stimati "notando" (*sic!*) la differenza tra beneficiari e non-beneficiari. Nessuna traccia della corposa letteratura che chiarisce sotto quali condizioni stringenti questo confronto *identifica* un effetto. Apprendiamo inoltre che la regressione ha lo scopo specifico di stabilire se i legami causali tra l'intervento e i suoi effetti sono "statisticamente significativi": si mescolano le problematiche dell'*identificazione* con quelle proprie della *stima*.

Secondo esempio. Tra i 5 scopi della valutazione visti nel grafico più sopra, un ruolo per il metodo sperimentale lo si trova in corrispondenza di quella che *The GUIDE* definisce come *knowledge production*: vediamo quale presentazione viene fatta dei "metodi e tecniche" a supporto di tale "produzione di conoscenza".

Typically, for knowledge production purposes, evaluators will want to answer the question, what works? From a positivist perspective, this would be an area where experimental methods are seen as relevant. However, the diverse and bottom-up nature of socio-economic interventions, the way these are combined in particular configurations and the different localities and contexts where programmes take place, makes traditional experiments difficult to apply except in very unusual circumstances. It is for that reason that realist thinking, with its emphasis on the influence of context on outcomes, has become more common in these kinds of evaluations. Here the more complex question is asked: what works, for whom, how and in what circumstances? Methods and techniques suitable for this will generally involve comparison between different cases selected to demonstrate alternative interventions and alternative contexts. Such comparisons may be based on case studies, data-bases that structure intervention/outcome/context configurations or a range of other techniques that are

¹³ *En passant*, notiamo come quel "netto" riferito ad "effetto" sia ridondante. Un effetto o è netto o non è un effetto. "Effetto lordo" è un ossimoro: per "effetto lordo" si intende in realtà "cambiamento osservato in concomitanza con...".

able to capture and describe these different aspects of socio-economic development.
(pag. 116).

Questo paragrafo è innanzitutto vittima della forma di miopia (a cui si è già fatto cenno) molto diffusa nella letteratura valutativa: quella di ignorare l'esistenza dei metodi non-sperimentali e concentrare la critica sul solo metodo sperimentale. Quindi il fatto che i metodi sperimentali siano "difficili da applicare" (a causa, si afferma senza minimamente chiarire, della natura *bottom-up* degli interventi socio-economici) porta a rigettare tutta la *positivist perspective* (cioè, decodificando, qualsiasi approccio quantitativo).

Il paragrafo fa poi eco alla critica "realista" dell'uso di esperimenti per valutare "*what works*", senza però approfondire tale critica. L'enfasi che i *realist* pongono "sull'influenza del contesto sui risultati" non è certo preclusa dall'uso di un approccio controfattuale, sperimentale o meno che sia. Paradossalmente, la critica più proficua e profonda dei *realist* al metodo sperimentale (in realtà, a tutto l'approccio controfattuale), quella che si può riassumere nell'affermazione "*experimentalists have pursued too single-mindedly the question of whether a program works at the expense of knowing why it works*"¹⁴, qui non viene neppure menzionata. Come non viene mai ripresa la tesi, sostenuta dai *realist*, che i "meccanismi" dovrebbero stare al centro della valutazione degli effetti. Evidentemente non è la prospettiva realista in sé che realmente interessa, ma solo il fatto che si contrapponga al metodo sperimentale.

Terzo esempio. Nella sezione 4.4 della *GUIDE* i metodi e le tecniche sono riclassificati a seconda del punto in cui sono utilizzati all'interno del "ciclo di *policy*". Ovviamente a noi interessano i metodi da utilizzarsi al "*conclusion/result stage*", dove si misurano risultati e "impatti".

(...)

Conclusions/Results: Outcomes and impacts

Policy makers for accountability reasons and key stakeholders, because of their own needs and commitments, look to evaluation to provide information on outcomes and impacts at the end of a programme cycle. Evaluation methods will seek to compare what has been achieved with what was intended and endpoints with baselines. A broad range of techniques can be deployed including:

- *Surveys of intended beneficiaries,*
- *Econometric or statistical models to demonstrate changes in economic performance compared with predicted results (perhaps by comparing trends in a development setting with other settings and using models developed at the beginning of a development cycle), and*
- *Indicators based on contextual data or administrative data provided by public authorities.* (pag. 120)

Qui la confusione regna sovrana: quella che era *knowledge production* ora lascia il posto all'*accountability*; la valutazione intende confrontare le realizzazioni con le intenzioni e i punti di arrivo con le condizioni iniziali, che sono due confronti dal significato molto diverso tra loro, nessuno dei quali *rivela* alcun effetto o impatto; i modelli econometrici o statistici servono a *dimostrare* cambiamenti in confronto ai risultati previsti (non sapevamo che i modelli econometrici avessero questo scopo!); l'elenco delle tecniche si chiude con un sibillino accenno agli indicatori.

¹⁴ Ray Pawson e Nick Tilley, *Realistic Evaluation*, London: Sage, 1997, pag. XV. A proposito della critica "realista" all'approccio controfattuale, va notato un punto fondamentale: i due approcci non sono incompatibili, per il semplice fatto che rispondono a esigenze conoscitive diverse. L'approccio controfattuale indaga *the effects of causes*, gli effetti di cause, mentre i realisti sono interessati alle cause di effetti, *the causes of effects*. Nel primo caso si vuole capire se un certo stimolo ha prodotto un effetto (desiderato o meno), nel secondo caso si vuole capire quali meccanismi causali hanno prodotto (o possono produrre) certi effetti (osservati o auspicati). Chi scrive è convinto che entrambi gli obiettivi conoscitivi siano importanti e che vadano perseguiti entrambi, riconoscendo la diversità degli strumenti analitici necessari a perseguirli: di sicuro il metodo sperimentale non serve a capire *the causes of effects*.

Nessun riferimento all'analisi controfattuale, che pure dovrebbe avere qualche ruolo nella fase conclusiva di un intervento, quando cioè si valutano "risultati e impatti". Se non qui, dove altrimenti? Non certo nella fase di disegno della *policy*, né quando si tenta di migliorarne l'implementazione.

Quarto esempio. Nella sezione dedicata agli indicatori apprendiamo che:

An indicator can be defined as the measurement of an objective to be met, a resource mobilised, an effect obtained, a gauge of quality or a context variable (pag. 127)

È chiaro, dunque: con gli indicatori si fa di tutto e di più, compreso "misurare effetti". Perché preoccuparsi di essere rigorosi e soprattutto cauti nello stimare gli effetti di un intervento, nel ricostruire nel modo più plausibile la situazione controfattuale, quando basta un semplice indicatore? L'utilità sovrana e soverchia degli indicatori è ripresa nel passaggio seguente:

The use of indicators in evaluation

Indicators serve a number of useful roles in evaluation. Their use is common with respect to programme evaluation, particularly where objectives are expressed in clear operational terms. The use of indicators normally forms part of an evaluation. The information they provide needs to be carefully interpreted in the light of other evidence in order that evaluative conclusions can be drawn. Indicators have the potential to contribute to the evaluation of socio economic programmes in several ways:

- *The analysis of the indicators scores can be used to provide support for a rationale for intervention and resource allocation.*
- *Indicators can be used to compare inputs and outputs in order to measure efficiency.*
- *Indicators can be used to compare actual outcomes with expectations in order to assess effectiveness.*
- *Indicators can be used to compare inputs relative to impacts and hence allow the assessment of the value (value added) of policy, legislation or initiatives.*
- *Indicators can be used to identify what would have happened in the absence of the initiative, policy or legislation (the counterfactual). (pag.130)*

I primi quattro punti non dicono nulla di nuovo rispetto alla retorica dominante nella letteratura valutativa, secondo cui con gli indicatori si allocano risorse, si misura l'efficienza, l'efficacia, l'economicità, ... ma all'ultimo punto *The GUIDE* aggiunge qualcosa di nuovo: gli indicatori servirebbero anche a *identificare il controfattuale*. Con buona pace di tutti i metodologi che hanno speso decenni a sviluppare questo approccio: hanno perso tempo, bastava un indicatore, ci fa sapere *The GUIDE*. Paradossalmente, questa è l'unica occasione in cui *The GUIDE* utilizza il termine *counterfactual*. Totalmente a sproposito.

4. Manca il dubbio, ecco la differenza

Perché *The GUIDE*, e con essa buona parte dell'*establishment* valutativo europeo, parla tanto di effetti, impatti, del bisogno di quantificarli, della loro importanza per i *policy-maker*, e poi tratta con tanta superficialità uno dei fondamentali approcci metodologici sviluppati a questo scopo dalla comunità scientifica? Crediamo che una possibile chiave esplicativa si possa rintracciare in una frase apparentemente innocua che troviamo nelle prime pagine del documento:

From an early stage, socio-economic development programmes need to demonstrate results.
(pag. 8) [enfasi aggiunta]

Dunque, il vero obiettivo non è mettere in dubbio l'efficacia di ciò che si fa, verificandone empiricamente gli effetti, quanto piuttosto "dimostrare risultati": quindi non una funzione di apprendimento, nonostante tutti i richiami alla *knowledge production*, bensì una esigenza (legittima, peraltro) di "rendere conto". *To demonstrate results* non occorre certo l'approccio controfattuale, basta fondamentalmente raccontare, descrivere, documentare ciò che si è realizzato: Non ci sarebbe

però bisogno di ammantare il discorso di termini altisonanti, quali “*outcomes and impacts*”. Dietro l’uso evocativo di questi termini non sta veramente, secondo chi scrive, un intento conoscitivo, teso a falsificare l’ipotesi che gli interventi realizzati producano gli effetti desiderati. Ci sta invece una diffusa, quasi istintiva, *presunzione di efficacia*: tutto ciò che è stato realizzato come previsto è per definizione efficace, quindi è sufficiente *descrivere* gli effetti. Quindi bastano gli indicatori.

Manca però un elemento fondamentale della valutazione degli effetti delle politiche, cioè il dubbio, l’incertezza, l’intenzione di mettere in discussione ciò che si fa. Il *dubbio sull’efficacia* dell’intervento pubblico è infatti ciò che contraddistingue, fin dagli inizi, la *program evaluation* americana e la sua domanda “*does this program work?*”

Questa differenza di intenti conoscitivi tra le due tradizioni valutative aiuta a spiegare perché oltreoceano si sia investito molto nello sviluppo e nell’uso di metodi per stimare gli effetti delle politiche, siano essi sperimentali o non-sperimentali; e perché invece in Europa, e massimamente in Italia, domini incontrastato l’uso di indicatori, spesso privi di un serio fondamento metodologico, come nel caso degli “indicatori di impatto”.

Concludendo. Chi il dubbio non ce l’ha, non se lo può dare. E troverà sempre un indicatore per *rassicurarsi* di aver fatto tutto bene.

Riferimenti bibliografici

- Bezzi C. (2003), *Il disegno della ricerca valutativa*, Franco Angeli, Milano.
- Blundell R., Costa Dias M. (2002), “Alternative approaches to evaluation in empirical microeconomics”, *Portuguese Economic Journal*, Vol. 1, No. 2, 91-115.
- Campbell D., Stanley J. (1966), “*Experimental and Quasi-experimental Designs for Research*”, Rand McNally, Chicago.
- Greenberg D., Linksz D., Mandel M. (2003), *Social Experimentation and Public Policymaking*, The Urban Institute Press, Washington.
- Greenberg D., Shroder M. (1997), *Digest of Social Experiments*, The Urban Institute Press, Washington.
- Heckman J., LaLonde R., Smith J. (1999), “The Economics and Econometrics of Active Labor Market Programs”, in Ashenfelter O., Card D. (a cura di), *Handbook of Labor Economics*, 1865-2097, North-Holland, Amsterdam.
- Heckman J., Smith J. (1995), “Assessing the Case for Social Experiments” *Journal of Economic Perspectives*.
- Holland P. (1986), “Statistics and causal inference”, *Journal of the American Statistical Association*. 81: 945-960
- Martini A, Mo Costabella L. e Sisti M. (2006) “Valutare gli effetti delle politiche pubbliche: metodi e applicazioni al caso italiano”, di prossima pubblicazione in *Materiali Formez*.
- Orr L. (1999), *Social Experiments: Evaluating Public Programs with Experimental Methods*, Sage Publications, Beverly Hills.
- Pawson R., Tilley N. (1997), *Realistic Evaluation*, Sage Publications, London.
- Tavistock Institute, GHK e IRS (2003), *The Evaluation of Socio-Economic Development: The GUIDE*, disponibile su www.evaled.info