Approfondimenti

19

SPERIMENTARE POLITICHE SOCIALI INNOVATIVE

Manuale introduttivo

QUADERNI DELL'OSSERVATORIO

fondazione cariplo

SPERIMENTARE POLITICHE SOCIALI INNOVATIVE

Manuale introduttivo (aggiornamento del 25 luglio 2014)

European Commission, Employment and Social Affairs and Inclusion
Traduzione italiana a cura dell'Ufficio Osservatorio e valutazione della Fondazione Cariplo

Collana "Quaderni dell'Osservatorio" n. 19 Anno 2015

Questo quaderno é scaricabile dal sito www.fondazionecariplo.it/osservatorio

doi: 10.4460/2015quaderno19

INDICE

ABSTRACT

Il lancio della Social Business Initiative da parte della Commissione europea ha recentemente rilanciato il dibattito sulla rendicontazione sociale delle attività delle organizzazioni di terzo settore.

Il tema non è nuovo, dato che da tempo anche molte imprese a scopo di lucro adottano forme di "bilancio sociale", talvolta unificando la rendicontazione economica tradizionale e quella socio-ambientale in un resoconto integrato. Anche nel settore pubblico esiste da tempo un'ampia esperienza (poco diffusa in Italia) che prende il nome di "valutazione d'impatto", con metodologie finalizzate alla valutazione controfattuale dell'effetto (e quindi dell'efficacia e dell'efficienza) delle politiche pubbliche. Non va infine dimenticato il filone di ricerca sulle misure di benessere sociale "oltre il Pil" che ha attirato un particolare interesse nell'ultimo decennio. Al fine di approfondire la conoscenza sulla pluralità di approcci, metodologie e strumenti disponibili per la valutazione, l'Osservatorio della Fondazione pubblicherà nella collana dei Quaderni dell'Osservatorio alcuni contributi specifici.

Questo Quaderno è la traduzione, realizzata dall'Ufficio Osservatorio e Valutazione della Fondazione, della guida "Testing social policy innovation – Primer for the Training" curata da LSE Enterprise e pubblicata dalla Commissione Europea nel luglio del 2014 con l'obiettivo di aiutare i candidati al Programma EaSI a redigere proposte corredate da strumenti di valutazione adeguati. L'obiettivo del Quaderno è fornire un'introduzione alla logica e ai principali strumenti della valutazione controfattuale, utile in primo luogo per gli enti e le istituzioni che presentano progetti e proposte di contributo alla Fondazione. Si tratta di un documento che fornisce una descrizione breve ma accurata delle metodologie più idonee a realizzare studi di valutazione dell'impatto (o degli effetti) di un progetto o un programma. Oltre alla parte metodologica la guida contiene una bibliografia esaustiva e alcuni casi di applicazione delle tecniche di valutazione in campo sociale.

La versione originale di questa guida è stata preparata e curata con il coordinamento di LSE Enterprise.

PREFAZIONE

Promuovere politiche di innovazione sociale significa sviluppare nuove idee, servizi e modelli che aiutino ad affrontare le sfide dei sistemi di *welfare*, alimentino la fragile ripresa economica attualmente in corso e migliorino i risultati sociali e occupazionali nel medio e lungo periodo. Tali politiche implicano nuove modalità di organizzazione dei servizi e il coinvolgimento di risorse pubbliche, private e della società civile. Per contribuire a raggiungere gli obiettivi di Europa 2020 sono infatti fondamentali partenariati più stretti tra questo ampio spettro di attori.

In quanto strumento in grado di fornire soluzioni migliori e innovative alle sfide sociali, l'innovazione sociale è un elemento essenziale per costruire riforme strutturali degli Stati membri coerenti con l'approccio del Social Investment Package (SIP). Il SIP sottolinea infatti la necessità di integrare l'innovazione sociale nel processo decisionale collegandolo alle priorità sociali. Sottolinea inoltre la necessità di modernizzare il welfare, messo in crisi dai cambiamenti demografici e dalla perdurante crisi economico-finanziaria. La modernizzazione delle politiche sociali richiede che le decisioni di spesa siano prese utilizzando un approccio orientato a raggiungere risultati definiti a priori, oltre che una considerazione sistematica del ruolo che le politiche sociali svolgono nelle diverse fasi della vita delle persone.

Infine, il SIP pone una particolare attenzione al miglioramento della misurazione dei risultati sociali, in particolare, in termini di rendimento sociale. Ciò significa la necessità di garantire che le riforme politiche siano non solo fondate sull'evidenza empirica, ma anche orientate ai risultati.

In questo contesto, il ruolo dei *policy maker* è cruciale sia nel guidare il processo di riforma, selezionando le priorità della politica, sia per garantire la tenuta e la sostenibilità dei risultati. Per giocare questa funzione, i responsabili politici hanno bisogno di strumenti che consentano di valutare i risultati delle politiche (aumento di inclusione e di occupazione, riduzione del costo del servizio a parità di qualità, contributo all'economia...).

I responsabili delle politiche hanno a disposizione diversi metodi a seconda delle caratteristiche specifiche della politica da valutare. Questi metodi possono misurare i risultati delle politiche, sostenendo e orientando le decisioni. La Commissione

europea organizzerà una serie di sessioni di formazione sulle diverse metodologie disponibili. Questa guida fa parte del materiale di formazione e verrà rivista e aggiornata sulla base dell'esperienza nell'utilizzo pratico. Questa guida integra un'iniziativa analoga sulle valutazioni d'impatto controfattuali nell'ambito di progetti finanziati sul Fondo sociale europeo (FSE).

INTRODUZIONE

La valutazione d'impatto misura gli effetti di una politica. La misurazione favorisce l'implementazione della politica sociale, ponendo in risalto il legame tra l'intervento realizzato e gli obiettivi perseguiti. Permette inoltre che le politiche efficaci siano replicate, contribuisce al loro continuo miglioramento e a un adeguato *follow-up*. *Policy maker*, fornitori di servizi e ricercatori si impegnano congiuntamente per valutare l'impatto delle politiche future e mettere i risultati a disposizione di tutti gli attori coinvolti.

Diversi metodi possono essere utilizzati per misurare l'impatto di una politica e ottenere prove dell'efficacia delle riforme adottate. Questa guida si concentrerà sui metodi più comunemente utilizzati, tra cui: (1) studi controllati randomizzati (*Randomized Controlled Trial*); (2) differenza nelle differenze (*Difference In Differences*); (3) abbinamento statistico (*Statistical Matching*) e (4) confronto attorno al punto di discontinuità (*Regression Discontinuity Design*).

Qual è lo scopo di questa guida?

Questa guida è un supporto per i responsabili politici e i fornitori di servizi sociali che desiderano implementare innovazioni di politiche sociali e valutare l'impatto dei propri interventi. Affronta tre questioni importanti e correlate:

- Come valutare l'impatto di un intervento di politica sociale. Quali metodi sono applicabili e sotto quali assunzioni funzionano?
- Come progettare una valutazione d'impatto? Le decisioni più importanti relative alla valutazione d'impatto saranno effettuate nella fase di pianificazione. Un piano affrettato e superficiale rischia di generare domande interessanti ma che rimarrebbero senza risposta o con risposte inadeguate a causa di dati parziali o mancanti. Questa guida fa luce – fornendo esempi concreti - sulle decisioni critiche che devono essere prese nella fase iniziale e i relativi *trade-off*.
- Come valutare e diffondere i risultati, sulla base della loro affidabilità, trasferibilità e sostenibilità. Come utilizzare la conoscenza acquisita perché supporti l'ottimizzazione delle riforme in corso, ispirando nuovo cambiamento e la costruzione di ulteriori conoscenze. Si tratta di partecipare a una comunità di *policy maker* che costruisce e condivide l'esperienza attraversando i confini settoriali.

Gli esempi nella seconda parte di questa guida, illustreranno il ruolo delle diverse metodologie nel sostenere e agevolare l'attuazione delle riforme progettate.

A chi è rivolta questa guida?

Questa Guida è destinata a sostenere, a livello nazionale, regionale e locale:

- i *policy maker*; gli attori coinvolti nel processo di costruzione delle politiche attraverso i programmi, le normative o il dialogo sociale. In particolare quelli tra loro che cercano di generare e/o utilizzare "evidenza empirica" su ciò che funziona (*what works*) per l'innovazione della politica sociale.
- i fornitori di servizi sociali; cioè le organizzazioni che erogano servizi sociali, che siano enti pubblici, organizzazioni *nonprofit* o imprese commerciali. Questa pubblicazione è particolarmente rilevante per coloro che cercano di valutare i risultati sociali dei loro programmi o delle loro politiche in modo credibile.

FASE 1

FASE 1 - DEFINIZIONE DI POLITICHE E INTERVENTI

Politiche sociali

Le politiche sociali sono costituite da diversi interventi interconnessi che affrontano problemi sociali. I singoli interventi che compongono una politica non possono essere considerati separatamente e il loro impatto dipenderà sia dalla loro interazione, sia da quella con le altre politiche (fiscali, finanziarie, ambientali, etc.). I singoli interventi possono essere valutati separatamente in modo affidabile, ma questa valutazione può riguardare anche interi programmi. Analisi individuale e complessiva si sostengono vicendevolmente e consentono una migliore comprensione, considerando contemporaneamente "la foresta e gli alberi".

Innovazione della politica sociale

Il concetto di innovazione politica è stato promosso dalla Commissione europea nel quadro dell'attuazione del Pacchetto di Investimenti Sociali (SIP). Si riferisce ad approcci d'investimento sociale che generano rendimenti sia economici che sociali ed è legato al processo di riforma dei sistemi di protezione sociale e di erogazione di servizi sociali attraverso innovative riforme di sistema.

Interventi di politica sociale

Un intervento è un'azione intrapresa per risolvere un problema. Nel campo della ricerca medica un intervento è un trattamento somministrato con lo scopo di migliorare un disturbo di salute. La relativa semplicità del trattamento medico lo rende facilmente replicabile; questo spiega in parte il motivo per cui l'idea stessa della sperimentazione sia emersa dal contesto medico e il metodo utilizzato risulti così convincente in tale ambito. Gli interventi di politica sociale, d'altra parte, hanno obiettivi diversi e forse meno focalizzati. Come accennato nel SIP, i sistemi di *welfare* svolgono

tre funzioni: l'investimento sociale, la protezione sociale e la stabilizzazione dell'economia. Per valutare gli interventi di politica sociale, potremmo quindi aver bisogno di combinare diversi metodi.

Selezione di un intervento, di un programma o di una politica rilevante

Mentre la strategia Europa 2020 e SIP offrono indicazioni chiare sulle priorità politiche, è molto importante identificare accuratamente gli interventi più rilevanti da valutare. Ad esempio, non hanno un grande valore aggiunto né la valutazione di interventi che interessano un numero molto limitato di persone né quella di una politica i cui effetti sono già stati solidamente provati. Sono invece ottimi candidati per la valutazione dell'impatto le modifiche del sistema di protezione sociale, gli interventi pilota innovativi o dimostrativi così come tutti gli interventi le cui conclusioni possono avere grande rilevanza per la più ampia comunità dei *policy maker*.

Mentre se ne valuta il potenziale impatto, è importante tenere a mente quali caratteristiche degli interventi possono essere testate in maniera attendibile; questa guida è pensata proprio per questo scopo. In ultima analisi, la decisione di testare un programma "poggerà su due gambe": la sua rilevanza e la sua fattibilità.

Valutare interi programmi e politiche

Una politica può essere valutata a diversi livelli, da quello 'macro' a quello 'micro'. Il livello appropriato dipende dalle esigenze del *policy maker*. C'è un *trade-off* tra la robustezza delle stime e la rilevanza politica concreta dell'intervento da valutare: un singolo intervento è infatti valutabile in modo assai più rigoroso di una politica più ampia che ha, però, probabilmente una rilevanza politica maggiore.

Si potrebbe quindi in primo luogo cercare di valutare l'impatto di un programma nel suo insieme con l'obiettivo di comprendere se questo, con tutte le sue componenti, ha creato una differenza sostanziale per i suoi beneficiari. Valutare un intero programma potrebbe creare però il cosiddetto problema della "scatola nera", a causa del quale il decisore pubblico e i ricercatori potrebbero non riuscire a distinguere il tutto dalle sue parti. Infatti, anche se il programma nel suo insieme mostrasse l'assenza di effetti significativi, tale risultato potrebbe essere messo in discussione, perché non sarebbero distinguibili il caso in cui nessun intervento ha avuto un effetto da quello in cui gli effetti dei diversi interventi potrebbero essersi annullati a vicenda a causa della loro interazione.

Valutare un intero programma ha anche implicazioni sugli strumenti metodologici disponibili. Ad esempio l'uso dell'analisi controfattuale, per sua natura, non è possibile in questo caso. Tuttavia, le limitazioni metodologiche potrebbero essere talvolta controbilanciate da vantaggi nella tempestività delle indicazioni che si potrebbero ottenere.

Valutare singoli interventi

A un livello più basso, può essere interessante testare ogni intervento separatamente. Confrontando l'impatto di ogni intervento, i decisori politici possono identificare l'al-

ternativa più efficace per raggiungere un determinato obiettivo. È importante notare che il costo della valutazione deve essere attentamente preso in considerazione al momento della progettazione dell'intervento. Tuttavia, la valutazione contemporanea di diverse ipotesi può produrre risultati più completi. Testare diverse ipotesi in un'unica valutazione può dunque risultare molto più efficiente in termini di tempo e risorse utilizzate.

Esempio: il supporto intensivo alla ricerca del lavoro (Intensive Job-Counselling)

Nel 2007, un gruppo di ricercatori ha valutato l'impatto di un programma di *counselling* intensivo fornito dai servizi per l'impiego francesi pubblici e privati. In primo luogo è stato misurato l'impatto di un aumento dell'intensità del servizio erogato dall'ente pubblico, e poi confrontato l'efficacia relativa del servizio di *counselling* intensivo fornito dal collocamento pubblico e da quello privato[1].

Gli interventi devono essere precisi e accuratamente definiti; estensioni o nuovi elementi del protocollo adottati in fase di applicazione possono determinare risultati molto diversi rispetto a quelli osservati durante la fase di valutazione.

Ulteriori informazioni

Glennerster R., Takavarasha K. (2013), *Running Randomized Evaluations: A Practical Guide*. Princeton University Press. pp. 8-12 e pp.73-77.

Morris S., Greenberg D., Riccio J., Mittra B., Green H., Lissenburgh S., Blundell R. (2004), *Designing a Demonstration Project*. London: Cabinet Office. Chapter 1. (www.civilservice.gov.uk).

National Audit Office (2011), *Auditing Behaviour Change*. pp.11-12. (www.nao.org.uk).

OECD, *Labour market programmes: coverage and classification*. (www.oecd.org).

1 Si veda l'Esempio 2 nella II Parte.

FASE 2

FASE 2 - SPECIFICARE UNA TEORIA DEL CAMBIAMENTO

Una teoria del cambiamento (*Theory of Change* - TOC)[2] sta a una nuova politica come le fondamenta stanno alla struttura di un edificio. La sezione che segue fornisce una rapida descrizione di questo approccio. Ulteriori informazioni sulle tappe più importanti della teoria del cambiamento sono riportate nelle sezioni successive.

Una mappa per il risultato desiderato

L'enfasi sulla progettazione deriva da un'osservazione più volte ripetuta dai ricercatori e dalle persone che a vario titolo si occupano di politiche pubbliche: molte importanti domande di valutazione rimangono infatti senza una risposta soddisfacente a causa di una progettazione superficiale. Sebbene la progettazione perfetta non esista, è possibile utilizzare una serie di passaggi in modo che l'energia utilizzata per lo sviluppo e lo svolgimento di una valutazione d'impatto produca i risultati che merita. Questi passaggi sono stati integrati in un unico quadro noto come 'teoria del cambiamento'.

La teoria del cambiamento è stata definita come "la descrizione di una sequenza di eventi che dovrebbe portare a un particolare risultato desiderato"[3]. Si tratta della catena causale che collega le risorse alle attività, le attività alle realizzazioni (*output*), le realizzazioni ai risultati (*outcome*) e i risultati ai cambiamenti (*impact*).

Una buona teoria del cambiamento utilizza sei diversi passaggi:

1. Bisogni: è la valutazione dei problemi sofferti dalla popolazione *target*.
2. *Input*: sono le risorse che saranno utilizzate per la realizzazione dell'intervento. Queste comprendono il tempo impiegato dagli implementatori e dai valutatori del progetto e i costi sostenuti per la sua attuazione (ad esempio per l'acquisto di beni e servizi). La domanda cruciale è: fino a che punto queste risorse consentiranno l'erogazione dell'intervento?

2 Si noti che la teoria del cambiamento è talvolta definita come "teoria del programma", "modello dei risultati", "logica di intervento", "quadro logico".

3 Rick Davies, April 2012: Post di un Blog sui criteri per giudicare la valutabilità di una teoria del cambiamento: "Criteria for assessing the evaluability of Theories of Change".

3. *Output:* è ciò che è stato erogato. Può consistere in una trasmissione di informazioni, nella fornitura di un sussidio o di un servizio, etc. La domanda chiave è: in che misura l'intervento è in grado di produrre le realizzazioni previste nel breve termine?
4. *Outcome*: sono i risultati osservabili che potrebbero essere raggiunti una volta che il servizio è stato erogato. I risultati nel settore delle politiche sociali di solito appaiono nel medio termine.
5. Impatto (*impact*): è il cambiamento nei risultati osservati attribuibile all'intervento sperimentato.
6. Infine, una teoria del cambiamento dovrebbe documentare le ipotesi (assunzioni) utilizzate per giustificare la catena causale.

Queste ipotesi devono essere supportate dalla ricerca e dalla consultazione degli *stakeholder*. Questo rafforzerà la plausibilità della teoria e la probabilità che i risultati dichiarati saranno effettivamente raggiunti.

Uno strumento essenziale nella gestione delle politiche sociali

Ci sono molti vantaggi nell'utilizzo di una teoria del cambiamento per sostenere l'innovazione della politica sociale. In primo luogo, una teoria del cambiamento aiuterà i responsabili politici a prendere decisioni migliori lungo tutto il ciclo di vita della politica. In una fase iniziale, sosterrà la formulazione di un'ipotesi chiara e verificabile su come si realizzerà il cambiamento. Ciò permetterà di migliorare non solo la capacità di rendere conto della politica (*accountability*), ma anche di rendere i risultati più credibili perché erano stati previsti. Durante l'implementazione, la teoria del cambiamento può essere utilizzata per controllare l'avanzamento del programma e mantenere la rotta, e anche come modello per la valutazione che produca indicatori misurabili del successo. Una volta che l'intervento è terminato, può infine essere aggiornata e utilizzata per documentare le lezioni apprese su quanto è realmente accaduto.

In secondo luogo, una teoria del cambiamento è un potente strumento di comunicazione in grado di catturare la complessità di un'iniziativa e sostenerla con i finanziatori, i *policy maker* o gli altri organi di governo. Il contesto economico difficile, così come la forte pressione su governi e organizzazioni perché dimostrino l'efficacia della propria azione, significa che i *leader* sono sempre più selettivi quando sostengono progetti di ricerca. Una rappresentazione visiva del cambiamento che la politica si propone sulla comunità e del modo in cui questo si realizzerà dovrebbe però rassicurare sulla credibilità dell'iniziativa. In questo modo il processo di attuazione e di valutazione può essere più trasparente, in modo che tutti sappiano cosa sta succedendo e perché.

Farlo bene

Una teoria del cambiamento è il risultato di due processi paralleli e simultanei che implicano ricerca e partecipazione. Il processo di ricerca mira a generare i dati alla base del programma e a verificarne i presupposti. Le aspettative sul fatto che un nuovo intervento porterà al risultato desiderato sono spesso giustificate dalla nostra 'esperienza' o 'buon senso'. Nella misura del possibile, le valutazioni d'impatto dovrebbero evitare di fare affidamento su tali misure soggettive che sono altamente discutibili e non offrono alcuna garanzia del successo dell'intervento. Per essere veramente dimostrato (*evidence based*), il nesso di causalità tra l'intervento e il risul-

tato dovrebbe contare su una solida base scientifica. Un intervento efficace richiederà contributi di varie discipline: economia, sociologia, psicologia, scienze politiche, etc. È fondamentale che tali competenze siano coinvolte sin dalle fasi preliminari del progetto. Il processo partecipativo di solito comprende una serie di incontri con gli *stakeholder*. L'obiettivo è duplice: (a) ottenere un *feedback* sulle conclusioni e le implicazioni della ricerca preliminare; e (b) ottenere il pieno coinvolgimento delle parti interessate, che è un fattore di successo essenziale.

Tabella 1 – Un esempio di teoria del cambiamento: il programma Pathways to Work

La tabella seguente mostra la teoria implicita sottostante le riforma delle prestazioni di inabilità al lavoro, come il programma The British Pathways to Work. Essa mostra che per ottenere un posto (impatto), il beneficiari del programma devono in primo luogo cercare un lavoro (*outcome*). I beneficiari vengono quindi invitati o incentivati a farlo (*output*). Questa catena di causalità vale nella misura in cui le assunzione del programma formulate dai responsabili sono credibili. Qui, il collegamento tra il risultato atteso "i beneficiari cercano un lavoro" e l'impatto atteso "i beneficiari ottengono un posto di lavoro" è subordinato alla loro competitività sul mercato del lavoro.

Teoria del cambiamento	Descrizione del programma	Assunzioni/ipotesi
Obiettivo	Assicurare la sostenibilità del sistema di assicurazione contro l'inabilità al lavoro	
Impatto	I beneficiari ottengono posti di lavoro	I beneficiari sono competitivi sul mercato del lavoro
Risultati (*outcome*)	I beneficiari presentano domande di lavoro	▪ I beneficiari sono convinti dai *case manager;* ▪ Gli incentivi economici sono sufficienti
Realizzazioni (*output*)	▪ Interviste obbligatorie focalizzate sul lavoro; ▪ Erogazione di incentivi economici per tornare al lavoro; ▪ Sistemi volontari per migliorare la disponibilità al lavoro.	▪ I beneficiari rispettano gli impegni presi; ▪ I beneficiari partecipano ai sistemi volontari
Risorse *(input)*	▪ Linee guida per le interviste obbligatorie sul lavoro; ▪ Formazione dei *case manager*; ▪ Risorse finanziarie; ▪ Software.	Le risorse economiche, di personale e le attrezzature sono sufficienti
Bisogni	Forte aumento del numero dei destinatari dei sussidi, con rischio per la sostenibilità della misura. Tra le cause del problema: ▪ deterioramento delle opportunità offerte dal mercato del lavoro; ▪ la politica unisce prestazioni di invalidità generose con un sistema permissivo di screening e monitoraggio.	

Ulteriori informazioni

Anderson A. (2005), The community builder's approach to theory of change: A practical guide to theory and development. New York: The Aspen Institute Roundtable on Community Change.

Sito web del Center for Theory of Change: http://www.theoryofchange.org/

FASE 3

FASE 3 – DEFINIZIONE DI RISULTATI, INDICATORI E PIANI DI RACCOLTA DATI

Le valutazioni di impatto testano il risultato atteso di un intervento. Ma quali sono le caratteristiche di un risultato ben definito? Che tipo di metrica deve essere utilizzata? In quale momento il risultato dovrebbe essere misurato? Questo capitolo fornisce alcune indicazioni per prendere le decisioni migliori.

Privilegiare gli effetti intenzionali ...

Un intervento può avere due tipi di effetti: intenzionali e involontari. La progettazione di una valutazione mira essenzialmente a identificare e valutare i primi. Data la complessità dei meccanismi sociali e la portata limitata della maggior parte degli interventi di politica sociale, i decisori pubblici dovrebbero individuare come risultato il problema che sono più desiderosi di risolvere e focalizzare su di esso tutte le energie e le risorse disponibili.

Nel caso ci fossero buoni motivi per aspettarsi ulteriori effetti positivi, sarebbe comunque buona norma assegnare in modo chiaro le priorità, distinguendo il risultato primario da quello o quelli secondari.

Esempio: il progetto pilota "Job Retention e Rehabilitation"

Tra il 2004 e il 2006, il Dipartimento britannico per il lavoro e le pensioni ha condotto il progetto pilota *Job Retention e Rehabilitation* composto da una serie di interventi, di tipo sanitario e sul luogo di lavoro, finalizzati ad aiutare le persone con problemi di salute cronici a conservare il lavoro. Data la natura dell'intervento, si è deciso che l'esito primario del progetto dovesse essere la situazione occupazionale dei partecipanti. La loro situazione sanitaria, che avrebbe potuto a sua volta essere influenzata dall'intervento, è stata invece considerata un risultato secondario.

... mentre si cercano anche gli effetti indesiderati

Mantenere l'attenzione sulla finalità principale dell'intervento non significa necessariamente che debbano essere ignorati i suoi eventuali segnali o esiti inattesi. Alcuni di loro potrebbero essere rilevanti. La maggior parte delle valutazioni - se non tutte - generano scoperte fortuite che possono mettere alla prova ed estendere la nostra

comprensione dei meccanismi economici e sociali. Tali fenomeni sono naturali nella ricerca sociale. Essi vanno quindi registrati, riportati e discussi e potrebbero giustificare ulteriori ricerche.

Scegliere la metrica giusta

Una volta stabilite le priorità, è importante identificare la metrica che fornirà la stima più accurata dell'impatto dell'intervento. La sfida consiste nel garantire che la variabile osservata rifletta effettivamente il risultato che doveva essere misurato. In altre parole, che la valutazione abbia una validità di costrutto.

Questo a volte è abbastanza semplice. Le politiche del lavoro, ad esempio, hanno tutte lo stesso obiettivo: aumentare il numero di persone occupate. Così, una valutazione avrà lo scopo di confrontare il numero di persone che ha ottenuto un posto di lavoro nel gruppo di intervento e in quello di controllo. Sarà probabilmente necessario discutere sulla definizione di ciò che si qualifica come lavoro - per esempio, un numero minimo di ore settimanali e un numero minimo di settimane di lavoro - ma la misurazione oggettiva dei risultati non dovrebbe soffrire di ulteriori problemi.

Altri risultati sono invece più difficili da misurare in quanto fanno riferimento a nozioni più complesse, più difficili da catturare con un solo indicatore. È il caso ad esempio degli indicatori delle capacità cognitive o della qualità della vita, etc. Per misurare questi risultati è meglio utilizzare scale composite o indicatori *proxy*.

Esempio: il "Medicare Alzheimer's Disease Demonstration"

> Nella loro valutazione del programma di cure a lungo termine Medicare Alzheimer Disease Demonstration, Yordi e colleghi (1997) hanno stimato l'impatto dell'intervento sulla qualità della vita dei beneficiari utilizzando la scala IADL (*Lawton Instrumental Activities of Daily Living*). La scala valuta le abilità di vita indipendente con otto strumenti, tra cui la preparazione del cibo, pulizia della casa e del vestiario. La scala IADL è ancora comunemente utilizzata negli Stati Uniti per valutare l'impatto degli interventi in campo sanitario.

Le misure dei risultati primari di un intervento dovrebbero, ove possibile, essere oggettive. Misure *soft* (ad esempio, gli atteggiamenti, i cambiamenti comportamentali auto-riferiti, i contatti a un determinato sito web a seguito di una campagna pubblicitaria) dovrebbero essere invece utilizzate per rafforzare i risultati o nei casi in cui non sia possibile ricorrere a misure oggettive. Se del caso, è preferibile elencare i diversi indicatori di risultato già utilizzati in letteratura e discuterli con esperti della materia.

Dichiarare le aspettative

Nella misura del possibile, si dovrebbe cercare di utilizzare gli stessi indicatori di risultato già impiegati in precedenti valutazioni di interventi simili (anche all'estero). L'utilizzo dello stesso indicatore non solo renderà più facili le rassegne sistematiche della letteratura e le meta-valutazioni, ma aiuterà anche a stimare la dimensione dell'effetto del nuovo intervento. Gli interventi possono avere effetti positivi o negativi e questo effetto può essere grande o piccolo. Per le ragioni che vedremo più avanti, le valutazioni d'impatto dovrebbero avere sempre lo scopo di valutare gli interventi con effetto atteso positivo e abbastanza grande - anche se la grandezza dell'effetto

dipende strettamente dalle caratteristiche specifiche della politica e del contesto. Inoltre, la valutazione dovrebbe cercare di focalizzarsi su una misura variabile al variare dei risultati.

Ad esempio, se è plausibile immaginare che una variazione delle prestazioni assistenziali di invalidità inciderà soprattutto sul ritorno al lavoro di coloro che hanno un livello intermedio di disabilità, può convenire misurare solo il tasso di occupazione di questo gruppo specifico dopo l'intervento. Il confronto dei livelli occupazionali di tutte le persone con disabilità potrebbe infatti produrre un risultato meno netto (o una stima meno attendibile), anche se, in questo caso, un numero maggiore di persone sarebbe (potenzialmente) influenzata dall'intervento e motivata a rientrare nel mercato del lavoro.

Ciò è importante per ragioni politiche e analitiche. In primo luogo, impostare obiettivi ragionevoli dell'intervento e documentare il raggiungimento di questo risultato aumenterà il sostegno politico per il programma. I responsabili politici potranno più facilmente promuovere una riforma se saranno in grado di documentare che l'intervento potrebbe ridurre di una misura definita (ad es. il 2%) il tasso di disoccupazione della popolazione *target*. Allo stesso modo, l'argomento risulterà più incisivo se potranno anche mettere a confronto ipotesi alternative, ad esempio che l'opzione A può ridurre la disoccupazione tra 0 e 4% e l'opzione B tra il 2 e il 6%.

In secondo luogo, gli impatti più grandi sono più facili da stimare. Ad esempio, nel caso di uno studio controllato randomizzato (RCT), maggiore è l'effetto atteso dell'intervento e più probabile sarà la sua osservazione anche con campioni di ridotte dimensioni.

Se l'intervento si basa su una riforma già avviata, una rassegna delle valutazioni precedenti fornirà un'indicazione dell'entità del probabile impatto. I valutatori dovrebbero sempre svolgere una rassegna sistematica degli studi sulle riforme già avviate. Se le rassegne sistematiche non sono disponibili, potrebbero condurne una come parte della fase di progettazione.

Effetto minimo rilevabile

L'effetto di un intervento deve essere sostanziale. Ciò significa che deve essere:

1. sufficientemente grande per giustificare il costo dell'intervento; questa è una domanda per il decisore politico, che deve confrontare il costo dell'intervento con i suoi benefici. Ciò significa che deve avere almeno un "effetto economico minimo";
2. sufficientemente grande da essere rilevato in un campione di dimensioni ragionevoli, cioè da un numero ragionevole di persone coinvolte nella valutazione. In questo caso, ragionevole significa che non può essere maggiore della popolazione disponibile e non può essere tanto grande da rendere la valutazione troppo costosa. Ciò significa che l'effetto deve essere maggiore di un "effetto minimo rilevabile".

Analizzare in dettaglio l'impatto dell'intervento

Dimostrare che un nuovo intervento ha avuto successo (o meno) nella risoluzione di un problema considerando l'intero *target* dei beneficiari è un elemento molto pre-

zioso. Tuttavia, questo risultato potrebbe essere percepito come un'analisi grossolana dell'impatto dell'intervento. Quindi, oltre a presentare i risultati complessivi, di solito i rapporti di valutazione sono in grado di produrre un più ampio insieme di informazioni dettagliate sulle condizioni e le circostanze nelle quali l'intervento è più efficace. Un'analisi più raffinata può essere infatti molto utile per i decisori pubblici.

Pianificare la raccolta dei dati: quando misurare l'impatto

Una valutazione d'impatto mira sostanzialmente a verificare che la differenza tra gli effetti dei due (o più) interventi sia tanto grande da non poter essere attribuita al caso. L'effetto di ogni intervento è calcolato come differenza tra il valore della variabile risultato (*outcome*) misurato prima e quello rilevato dopo l'esecuzione dell'intervento, a parità di tutte le altre variabili (*ceteris paribus*; vedi sezione 5).

La misura effettuata prima della realizzazione dell'intervento è chiamata *baseline*. La misura a fine progetto dovrebbe essere rilevata nel momento in cui il progetto dovrebbe aver prodotto il suo effetto (impatto). Il momento esatto dipende da una serie di fattori, in primo luogo il tipo di intervento: mentre alcune tipologie di azione, come ad esempio le campagne di informazione, hanno un effetto piuttosto immediato, altre potrebbero richiedere anni per produrre un effetto (è il caso ad esempio dei programmi di istruzione). Inoltre, gli interventi complessi spesso richiedono una fase pilota per lasciare che il programma sia assimilato e consentire agli operatori sul campo di familiarizzare con i loro nuovi compiti. Secondo il Magenta Book[4], anche

4 Il Magenta Book, la guida alla valutazione consigliata dal governo britannico, definisce le migliori pratiche da seguire negli uffici pubblici. E': https://www.gov.uk.

se non c'è una durata prefissata per una valutazione d'impatto, di solito servono "almeno due o tre anni"[5]. Indipendentemente da quando avvenga la rilevazione finale, è comunque importante rispettare il protocollo di ricerca.

La misurazioni prima e dopo l'intervento sono quindi le due estremità e i capisaldi della valutazione di impatto. Un'altra buona idea è predisporre un monitoraggio durante la realizzazione del programma per verificarne l'andamento e testare le variabili risultato. Questa misura può essere vista come una 'prova generale' per la misura finale. Data l'importanza di queste indagini, è essenziale che ognuno svolga puntualmente il proprio ruolo. Se qualcosa va storto - come ad esempio un numero troppo elevato di abbandoni del programma - è sicuramente meglio scoprirlo in una rilevazione intermedia che in quella finale.

Si raccomanda che le valutazioni intermedie siano rese pubbliche, come tutti gli altri esiti di ricerca, una volta che la valutazione è giunta al termine. Da un punto di vista politico, si potrebbe essere tentati di interrompere la valutazione dopo la misurazione intermedia. Tuttavia si dovrebbe sempre resistere a questa tentazione, infatti, la maggior parte degli interventi di politica sociale ha effetti molto diversi dopo sei o dodici mesi. Siccome, in ultima analisi, quello che conta per i decisori pubblici è l'effetto a lungo termine, i risultati a breve termine dovrebbero sempre essere considerati con cautela.

La scelta del metodo più appropriato di raccolta dei dati

Il modo più economico per la raccolta dei dati è utilizzare quelli di fonte amministrativa generati autonomamente dai fornitori di servizi e/o dalle autorità pubbliche. Ad esempio, i fornitori di servizi di formazione professionale tengono dei registri dettagliati con dati relativi a ciascuno studente, il numero di ore/sessioni di formazione cui hanno partecipato, la data di uscita (di chi ha interrotto il percorso formativo), la ragione dell'interruzione (ad esempio, lo studente potrebbe aver trovato un posto di lavoro, o potrebbe aver rinunciato per motivi di salute). Tuttavia, l'esperienza insegna che in alcuni casi tali dati non sono sufficienti a rispondere a tutte le questioni sollevate dal programma. Al contrario, indagini nazionali su larga scala tendono a raccogliere dati molto dettagliati (ad esempio sul reddito, il numero di ore lavorate, etc.), ma solo per un piccolo sottoinsieme della popolazione. In questo caso l'informazione sarebbe disponibile solo per una frazione dei partecipanti. In ogni caso, una verifica dettagliata dei dati disponibili, della loro completezza e affidabilità è un criterio essenziale per la progettazione di una valutazione di impatto adeguata.

Il metodo più comune di raccolta dei dati per le valutazioni di impatto è attraverso i sondaggi. I sondaggi sono uno strumento molto flessibile: possono essere usati per ottenere informazioni sui partecipanti, per misurare le loro opinioni e atteggiamenti sul tema oggetto di intervento e per raccogliere i dati di *outcome* (per esempio lo stato di occupazione, le numero di ore lavorate, il reddito, etc). Tuttavia questo tipo di dati può comportare un rischio di affidabilità: i partecipanti potrebbero aver dimenticato quello che è successo in passato o potrebbero, più o meno intenzionalmente, fornire dati errati su informazioni riservate, come il loro reddito. Data l'importanza di tali

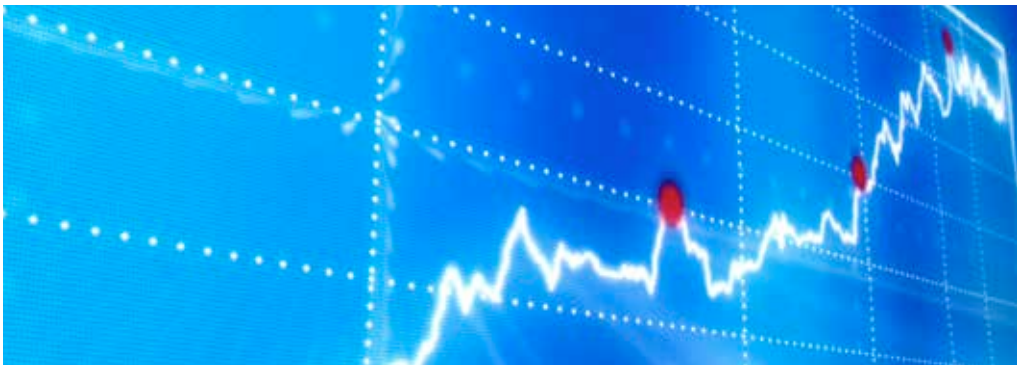
5 http://www.civilservice.gov.uk/wp-content/uploads/2011/09/chap_6_magenta_tcm6-8609.pdf

indagini, è importante progettare tutti i questionari con grande cura e con l'aiuto dei valutatori.

Ulteriori informazioni

Glennerster R., Takavarasha K. (2013), Running Randomized Evaluations: A Practical Guide. Princeton University Press. pp. 8-12 e pp.73-77.

Morris S., Greenberg D., Riccio J., Mittra B., Green H., Lissenburgh S., Blundell R. (2004), Designing a Demonstration Project. London: Cabinet Office. Chapter 1. (www.civilservice.gov.uk).



FASE 4

FASE 4 - LA STIMA DEL CONTROFATTUALE

Le valutazioni di impatto cercano di stimare il valore intrinseco delle politiche pubbliche. Ci sono molte ragioni per cui un programma potrebbe essere percepito come un successo, anche se non ha avuto alcun impatto reale o viceversa. Per esempio, l'attuazione del programma potrebbe aver coinciso con condizioni economiche favorevoli, nel qual caso la situazione sarebbe migliorata anche senza il nuovo programma. O, confrontando gli effetti tra due gruppi, si potrebbe riscontrare che i beneficiari del nuovo intervento siano un po' diversi dai membri del gruppo di controllo, con ciò aumentando (o abbassando) artificialmente gli effetti rilevati.

Per tener conto degli effetti che non hanno nulla a che fare con l'intervento, le valutazioni di impatto misurano l'esito rispetto a una stima di ciò che sarebbe accaduto in assenza dell'intervento. Questa stima è nota come controfattuale[6].

Ci sono diversi modi di stimare il controfattuale. Questo capitolo descrive brevemente la differenza tra controfattuale a livello individuale e controfattuale a livello di popolazione, presentando le tecniche utilizzate più frequentemente. Ciascuno di questi metodi si basa su una o più ipotesi più o meno credibili a seconda del contesto della valutazione e dei dati disponibili. È importante che sia il valutatore, sia il *policy maker* siano consapevoli di questi presupposti e interpretino i risultati ottenuti con le riserve del caso.

Controfattuale a livello individuale e a livello di popolazione

Mentre alcune riforme mirano a modificare i risultati a livello di popolazione, altri cercano di influenzare il comportamento dei partecipanti a un programma (i risultati a livello individuale). É molto difficile misurare i risultati a livello di popolazione e risultati a livello individuale contemporaneamente.

6 Si veda a proposito; ESF Guide on Counterfactual Evaluation: Design and Commissioning of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2012.

Per stimare l'impatto di un intervento in base ai risultati ottenuti a livello individuale (ad esempio, la partecipazione al mercato del lavoro, l'utile netto, la durata delle prestazioni) occorre costruire controfattuali a livello individuale. Tali controfattuali prevedono che i risultati sui beneficiari di un nuovo intervento siano confrontati con i beneficiari delle disposizioni in vigore (controfattuale).

Controfattuali costruiti a livello individuale possono essere utilizzati anche per stimare l'impatto di diversi aspetti della riforma e in questo modo identificare gli elementi più efficienti. Ciò è molto importante da un punto di vista politico, perché concentrare la spesa sulle opzioni che hanno un maggiore impatto sui risultati desiderati consente di razionalizzare la spesa pubblica. I controfattuali a livello individuale aiutano inoltre a stimare l'eterogeneità dell'impatto su gruppi di popolazione differenti consentendo di costruire azioni su misura. Questo può essere molto importante per alcune politiche come, ad esempio, quelle rivolte all'attivazione dei percettori di sussidi, come il reddito di cittadinanza. Infatti, oltre a costituire un gruppo molto eterogeneo, queste persone incontrano generalmente maggiori difficoltà di impiego (ad esempio rispetto ai beneficiari di un sussidio di disoccupazione).

Al fine di valutare l'impatto della riforma sui beneficiari di un nuovo servizio, si ha la necessità di costruire controfattuali a livello di popolazione. In questo caso, si confrontano due gruppi simili: un gruppo accede al nuovo regime, mentre l'altro continua a beneficiare del regime esistente. La differenza nei tassi di partecipazione o iscrizione e nelle caratteristiche individuali dei beneficiari dei due regimi può offrire una misura della variazione dei flussi in entrata e della composizione della popolazione indotta dal nuovo servizio. Tuttavia, questo tipo di analisi può essere difficile da svolgere, in particolare se la fornitura di servizi è frammentata tra organizzazioni o reparti differenti.

Metodo 1 - Studio randomizzato controllato (Randomized Controlled Trial - RCT)

La credibilità di una valutazione di impatto dipende dal grado di somiglianza tra il gruppo di controllo e il gruppo di intervento, in termini di caratteristiche sia osservabili che non osservabili. L'assegnazione casuale al gruppo di intervento e a quello di controllo costituisce il metodo più affidabile; se il campione è infatti sufficientemente grande, si ha la garanzia che i due gruppi abbiano le stesse caratteristiche[7].

Spesso le ragioni che determinano la partecipazione a un intervento sono correlate con i risultati. Ad esempio, è molto probabile che saranno proprio le persone più motivate quelle che più tenderanno a iscriversi a un programma di consulenza per la ricerca di un'occupazione. La motivazione a sua volta può essere correlata con la probabilità di trovare un lavoro. Confrontando semplicemente partecipanti e non partecipanti a un programma di questo tipo si otterrà quindi una misura sovrastimata

7 Distinguiamo tra metodi (sperimentali) prospettici e metodi retrospettivi (quasi-sperimentali). In quest'ultimo caso non c'è un'assegnazione casuale ma una manipolazione della variabile indipendente in modo da creare un gruppo di confronto statistico e/o intertemporale (*reflexive*). Quando non è pratico o etico assegnare casualmente i partecipanti a un programma (ad esempio, maschio o femmina o a specifiche categorie, etc...) si utilizzano dei "quasi-esperimenti" che sono progettati per massimizzare la validità interna, anche se questa tenderà a rimanere comunque inferiore a quella di uno studio randomizzato. Il resto di questo capitolo fornisce una panoramica dei principali metodi disponibili.

dell'impatto dell'intervento; se i partecipanti sono anche (mediamente) i più motivati, essi avrebbero comunque una maggiore probabilità di trovare un lavoro, anche in assenza del programma. È comunque possibile anche il contrario: se un programma di consulenza per la ricerca di un'occupazione fosse disponibile solo per i disoccupati non qualificati di lungo periodo, il confronto tra partecipanti e non partecipanti ci fornirebbe una misura sottostimata dell'impatto, in quanto i partecipanti se la sarebbero cavata relativamente peggio dei non partecipanti anche in assenza del programma.

Questi esempi dimostrano l'importanza dei fattori di selezione. L'assegnazione casuale da un campione sufficientemente ampio di soggetti a un gruppo di intervento e a un gruppo di controllo ci aiuta a tenere sotto controllo i problemi di selezione, in quanto garantisce che i gruppi siano statisticamente identici sia sulle caratteristiche osservabili sia su quelle non osservabili. In questo modo, l'unica differenza tra i due gruppi è la partecipazione o meno all'intervento (almeno temporaneamente). Così facendo è più facile stabilire relazioni causali tra l'intervento e la differenza dei risultati osservata tra partecipanti e non partecipanti.

Gli studi randomizzati controllati sono considerati un metodo rigoroso per la costruzione di un controfattuale valido. Tuttavia, l'applicazione di questa tecnica per la valutazione di programmi in campo sociale richiede tempo e risorse, e soprattutto, la necessità che il disegno di valutazione sia progettato prima dell'implementazione dell'intervento.

Gli studi randomizzati e controllati non sono sempre applicabili. Gli interventi sociali sono spesso vincolati da leggi e procedure amministrative. Ad esempio, la costituzione di un paese potrebbe vietare di rivolgere esclusivamente a un sottoinsieme della popolazione un intervento che comporta la riduzione o l'aumento dei benefici; alcuni interventi potrebbero applicarsi a intere comunità, come ad esempio, quelli che promuovono economie parallele di mutuo-aiuto. In generale, tuttavia, un piano di riforma prevede solitamente alcune misure che possono essere testate con studi randomizzati e altre misure che richiedono altri metodi.

Esempio: la sanità integrata negli Stati Uniti

Uno studio americano pubblicato nel 2002 ha riguardato un gruppo di anziani invalidi, beneficiari di servizi di assistenza domiciliare, esposti al rischio di utilizzare una quantità troppo elevata di servizi per malati acuti. Metà dei pazienti sono stati quindi assegnati in modo casuale a un'infermiera (*Nurse Care Manager*) che aveva il compito di migliorare il collegamento tra i servizi per acuti e quelli di assistenza a lungo termine già utilizzati dai pazienti coinvolti nel programma. Lo scopo dell'intervento era ridurre i ricoveri ospedalieri.

Nonostante qualche piccola differenza nell'uso dei servizi sanitari e nel costo degli stessi tra il gruppo sperimentale e quello di controllo, gli autori hanno concluso che non vi erano differenze tra i due gruppi in nessuna delle variabili risultato (*outcome*) esaminate durante i 18 mesi di sperimentazione. Gli sforzi per integrare i sistemi di cura dell'acuzie e di assistenza a lungo termine si sono dimostrati più difficili del previsto. L'intervento, che ha tentato di raggiungere l'obiettivo attraverso il ricorso ai servizi di un *case manager*, senza incentivi finanziari o regolamentari, non si è rivelato in grado di produrre un cambiamento significativo per i pazienti che ne hanno beneficiato. Il programma è stato anche influenzato da vari cambiamenti organizzativi, come ad esempio quelli nella gestione degli ospedali coinvolti negli studi, con ripercussioni sul loro modo di comunicare con il *case manager* (Applebaum *et al.*, 2002).

Quando la randomizzazione è utilizzata per assegnare partecipanti ai gruppi di intervento e di controllo, vi è un'alta probabilità che i due gruppi siano identici. Questa ipotesi può essere testata empiricamente attraverso una prova di bilanciamento. I ricercatori possono misurare la distribuzione delle caratteristiche nei gruppi coinvolti nella valutazione in modo da verificare che non ci siano differenze significative nelle variabili chiave che potrebbero influenzare i risultati. La mancanza di bilanciamento potrebbe comunque verificarsi anche se il processo di assegnazione casuale è stato effettuato correttamente, ma questo rischio si minimizza al crescere delle dimensioni del campione.

Per questo test valgono quindi gli stessi *caveat* che si applicano alla tecnica dell'abbinamento statistico (vedi sotto): non c'è mai la garanzia che il test includa tutte le variabili rilevanti e, conseguentemente, la composizione dei due gruppi può essere distorta per effetto di una variabile non osservabile. È comunque rassicurante che non vi siano differenze significative nelle variabili osservabili.

Uso tipico degli studi randomizzati e controllati (RCT)

L'assegnazione casuale è un disegno valutativo valido quando sono soddisfatte le seguenti condizioni:

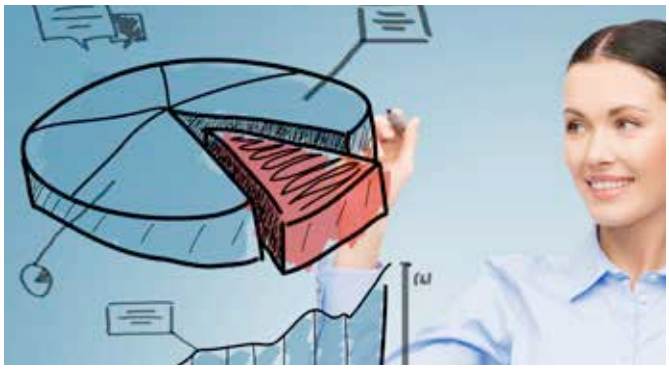
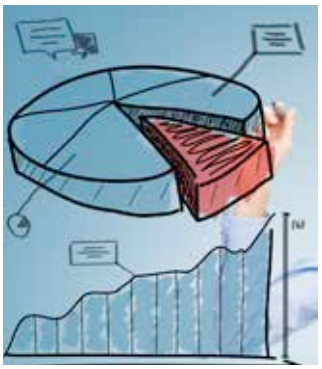
1. Gli aspetti etici collegati alla ricerca su soggetti umani sono stati ben considerati e non impediscono l'applicazione di diversi interventi a persone diverse. Ad esempio, nel caso non vi fossero vincoli di risorse, non sarebbe etico negare a qualcuno un intervento i cui benefici sono già stati documentati al solo fine di realizzare un esperimento.
2. La dimensione del campione è abbastanza grande. Nel caso non vi sia un numero adeguato di partecipanti al progetto pilota, potrebbero non esserci sufficienti osservazioni per rilevarne statisticamente l'impatto (anche in caso di successo).

Metodo 2 - Confronto attorno al punto di discontinuità (Regression Discontinuity Design - RDD)

Questo metodo può essere implementato in presenza di un chiaro criterio quantitativo di ammissibilità o valore soglia che separa un gruppo di soggetti ammessi all'intervento da un altro gruppo (controllo). Il metodo mette a confronto i soggetti ammessi all'intervento con punteggi appena superiori alla soglia prefissata con i non ammessi che hanno ricevuto punteggi appena sotto tale soglia.

A titolo di esempio, si supponga che a seguito di una nuova disposizione relativa agli assegni di invalidità, alle persone al di sotto di una certa soglia di invalidità siano ridotti i vantaggi e offerte misure attive per l'ingresso nel mercato del lavoro. Il metodo dovrebbe quindi confrontare disabili appena sopra la soglia (che godrebbero dei vecchi benefici) con quelli appena sotto tale livello (ai quali si applicherebbe la nuova disposizione).

Questo metodo si basa sul presupposto che i soggetti sono ammessi all'intervento con un criterio di selezione quantificato chiaramente e che i partecipanti non sono in grado di prevedere e manipolare i punteggi vicini al punto di *cut-off* (nell'esempio precedente, 'modificare il loro livello di disabilità'). Si assume inoltre che gli individui appena sotto e appena sopra la soglia non siano significativamente diversi. Questo di


solito comporta che il punteggio intorno la soglia sia un *continuum*. Potrebbero invece verificarsi differenze significative se, per esempio, la stessa soglia venisse utilizzata per fornire diversi servizi, nel qual caso i due gruppi (sopra e sotto la soglia) affronterebbero condizioni qualitativamente differenti anche senza considerare l'intervento.

La principale criticità di questo metodo è che misura l'effetto dell'intervento solo sulle persone che si trovano vicino alla soglia di ammissibilità. Se i decisori politici sono interessati a valutare l'impatto di una politica o di un programma su tutta la popolazione, questo metodo non è appropriato; i risultati, per esempio, non possono essere utilizzati per analizzare l'impatto di un'eventuale variazione (verso l'alto o il basso) del valore di soglia. Un altro problema pratico è decidere l'ampiezza dell'intervallo di punteggio utilizzato per determinare il campione. Da un lato, se l'intervallo è piccolo, i due gruppi (appena sopra e sotto) saranno simili, ma l'effetto sarà misurato su poche persone, con conseguente maggiore incertezza. D'altra parte, se l'intervallo è ampio e comprende molte persone, sarà idealmente possibile ottenere una stima più precisa dell'effetto, ma confrontando gruppi più differenti fra loro.

Esempio: la riforma dell'assicurazione sull'invalidità in Norvegia

Kostol e Mogstad (2014) hanno utilizzato questo metodo per valutare l'impatto di un cambiamento nella politica di erogazione degli incentivi a favore dei beneficiari di Indennità per l'inabilità al lavoro, in Norvegia. Gli individui divenuti beneficiari prima del 1 gennaio 2004 erano infatti stati esposti a norme più generose sulla possibilità di sommare sussidi e salari (nuovi incentivi al lavoro), rispetto ai beneficiari del periodo successivo. Gli autori hanno ipotizzato che i richiedenti subito prima e subito dopo tale data fossero molto simili fra loro, per cui le differenze nei risultati (tasso di occupazione mentre si gode del sussidio o tasso di uscita dal mercato del lavoro) potrebbero essere attribuite alla variazione delle norme sugli incentivi. Questo non è un confronto prima-dopo, perché tutti gli individui sono osservati simultaneamente nella stessa situazione di contesto. Gli autori hanno scoperto che gli incentivi economici inducono una parte sostanziale dei beneficiari a tornare al lavoro, ma solo la componente più giovane. Questo conferma l'affermazione che alcuni beneficiari della misura possono lavorare e che gli incentivi sono efficaci nell'incoraggiarli a farlo.

Uso tipico del confronto attorno al punto di discontinuità (RDD)

Uno studio si qualifica come tale se l'assegnazione al programma può essere basata su un'unica variabile di discriminazione. Le unità con punteggi pari o superiori / inferiori a un determinato valore soglia sono assegnate al gruppo sperimentale mentre le unità con punteggi dal lato opposto della soglia sono assegnate al gruppo di controllo. Tale variabile deve soddisfare tre criteri:

1. deve essere continua o ordinale, con un numero sufficiente di valori univoci. La variabile non dovrebbe mai essere basata su categorie non ordinali (come il genere).
2. la variabile utilizzata non deve generare confusione. Lo stesso valore di soglia non deve essere utilizzato per assegnare ai soggetti interventi diversi da quello in fase di test. Ad esempio, l'ISEE (Indicatore della Situazione Economica Equivalente[8]) non può essere la base di un confronto attorno al punto di discontinuità (RDD), perché lo stesso criterio è utilizzato per l'ammissibilità ad una più vasta gamma di servizi. Questa limitazione è necessaria per garantire che lo studio possa isolare gli effetti causali dell'intervento testato dagli effetti di altri interventi.
3. il valore della variabile di discriminazione non deve poter essere manipolato: per esempio, il reddito familiare può essere riportato in modo non fedele per rendere i cittadini ammissibili a un programma o a un'agevolazione.

Metodo 3 - Differenza nelle differenze (Difference In Differences)

Questo metodo confronta la variazione nei risultati nel tempo tra partecipanti e non partecipanti. Più specificamente, si misura la variazione dei risultati ottenuti dal gruppo di controllo per avere un'idea di ciò che sarebbe stato il cambiamento naturale in assenza del programma, e si segue la variazione del risultato nel gruppo dei beneficiari per ottenere una misura della somma tra cambiamento naturale e cambiamento causato dal programma. Sottraendo la differenza dei risultati del gruppo di controllo (cambiamento naturale) a quella del gruppo dei beneficiari, il valutatore può quindi ottenere una misura della variazione causata dal programma. Uno dei vantaggi di questo metodo è che fornisce una misura dell'impatto per l'intera popolazione, controllando al tempo stesso il cambiamento delle condizioni di contesto. Tuttavia, si basa sull'assunzione che le tendenze siano parallele. In altre parole, per determinare la differenza dei risultati attribuibile al programma, si assume che la dinamica spontanea sia la stessa tra partecipanti e non partecipanti al programma. Un modo per confermare la credibilità dell'ipotesi delle tendenze parallele è quello di verificare e confrontare i cambiamenti nei periodi precedenti all'avvio dell'intervento. Questo esercizio richiede molti periodi di dati prima dell'intervento, sia per il gruppo dei beneficiari dell'intervento, sia per il gruppo di controllo. È anche importante verificare che non vi siano fenomeni locali - indipendenti dal programma - che potrebbero influenzare le tendenze mentre l'intervento viene attuato. Per esempio, l'applicazione di un altro programma nella regione di intervento o di controllo, o uno *shock* esogeno relativo a uno solo dei due gruppi potrebbe influenzarne l'esito e distorcere la misura dell'impatto stimato del programma.

8 Si tratta dello strumento di valutazione, attraverso criteri unificati, della situazione economica di chi richiede prestazioni sociali agevolate o l'accesso a condizioni agevolate ai servizi di pubblica utilità (www.inps.it).

Esempio: il programma Pathways to Work

La valutazione inglese del programma Pathways to Work ha misurato gli effetti di tale intervento sul tasso di uscita dal regime degli assegni di invalidità al lavoro e sull'occupazione generata (Adam, Bozio e Emmerson, 2010). Questa riforma pilota prevedeva forti incentivi economici per tornare al lavoro, interventi di monitoraggio individuale (interviste obbligatorie) e di attivazione (sistemi di consulenza volontari già citati nel capitolo Fase 2). La riforma è stata sperimentalmente introdotta progressivamente nei vari distretti. I funzionari del Dipartimento per il Lavoro e le Pensioni (DWP) hanno selezionato i distretti pilota prima dell'intervento e hanno lasciato ai valutatori la scelta dei distretti di controllo più adeguati sulla base di un insieme di caratteristiche osservabili a livello aggregato. Essendo una misura con variazioni attese a livello di quartiere, il controfattuale è stato costruito a livello di popolazione. In ciascun distretto sono state quindi seguite le persone entrate nel programma prima e dopo l'inizio del pilota. La differenza di risultati riscontrati dopo l'attuazione della politica tra le aree pilota e quelle di controllo è stata interpretata come un effetto del programma Pathways to Work.

Uso tipico del metodo della differenza nelle differenze (DiD)

Nella sua versione più semplice, il metodo può essere utilizzato per valutare l'impatto di un intervento nel caso in cui si possa disporre di dati relativi a due periodi di tempo. Nel primo periodo - pre intervento - nessun individuo è esposto alla nuova politica. Nel secondo periodo - post-intervento - quelli assegnati al gruppo di intervento sono già stati esposti alla politica mentre quelli del gruppo di controllo ne sono rimasti esclusi. Versioni più generali di questo metodo consentono l'adozione progressiva del programma da parte della popolazione di riferimento. Per utilizzare questo metodo per identificare l'impatto di un intervento, si deve presumere che i due gruppi - in assenza dell'intervento - avrebbero sperimentato tendenze simili nelle variabili risultato. Argomenti a favore di questo metodo possono venire dalla disponibilità di informazioni relativi ai periodi precedenti l'inizio dell'intervento: se le tendenze osservate si sono dimostrate parallele in passato è probabile che abbiano continuato a rimanere tali anche nei periodi più recenti.

Metodo 4 - Abbinamento statistico (Statistical Matching)

L'abbinamento statistico è il nome attribuito collettivamente a varie tecniche statistiche che costruiscono un gruppo di controllo accoppiando ciascuno dei partecipanti ad un intervento con un soggetto analogo non partecipante, sulla base delle caratteristiche osservate. L'obiettivo è quello di accoppiare i partecipanti con i non partecipanti utilizzando il maggior numero di variabili possibili, al fine di garantire che l'unica differenza (o la differenza principale) tra i due gruppi sia la partecipazione all'intervento. I non partecipanti abbinati rappresentano il controfattuale.

Un abbinamento di successo richiede una ricerca preliminare approfondita, al fine di identificare le diverse variabili che potrebbero essere statisticamente correlate alla probabilità di partecipare al programma e al risultato di interesse. In aggiunta a questo è necessario un campione sufficientemente ampio in modo da creare corrispondenze sufficienti.

Questo metodo fornisce una stima dell'effetto di un intervento per tutti i partecipanti che siano stati abbinati con successo a un non partecipante e, se vi sono abbastanza

dati disponibili, può essere applicato anche se il programma è già terminato. Tuttavia, questo metodo si basa sul presupposto forte e non verificabile che tutte le caratteristiche di base si possano osservare e contabilizzare. In pratica, non c'è modo di escludere un'eventuale distorsione causata da variabili non osservabili che potrebbero influenzare sia la partecipazione dell'intervento sia il risultato finale.

Esempio: la riforma del Workfare in Argentina

Jalan e Ravallion (2003) hanno usato questo metodo per testare l'impatto sul reddito di una riforma del sistema di welfare in Argentina. In mancanza di dati di *baseline*, ed essendo la valutazione stata progettata dopo l'attuazione del programma, i ricercatori hanno optato per una tecnica di abbinamento statistico. Attraverso un sondaggio, hanno raccolto informazioni su circa 200 caratteristiche individuali in modo da poter abbinare ad ogni partecipante al programma il non partecipante più simile. Hanno quindi calcolato la differenza media di reddito tra i gruppi risultanti dall'abbinamento e verificato la robustezza dei risultati applicando le diverse tecniche disponibili. Tuttavia, i ricercatori non hanno potuto escludere le distorsioni provocate dalle variabili non osservabili.

Utilizzo tipico dell'abbinamento statistico (SM)

1. I fattori che influenzano la partecipazione al programma devono essere noti. Ciò può essere un problema nel caso di programmi innovativi che adottano nuovi metodi per assistere i clienti. In questo caso, potrebbe essere necessaria una ricerca preliminare per individuare i fattori che potrebbero influire sulle scelte di partecipazione.
2. Devono essere disponibili informazioni sulle scelte di ingresso al programma e sui risultati di interesse. Tali informazioni potrebbero mancare nel caso non fossero stati allocati tempi e/o risorse sufficienti alla loro raccolta prima dell'avvio del programma (ad esempio sui redditi e le carriere lavorative).
3. Non esistono variabili non osservate che potrebbero influenzare la partecipazione e i risultati d'interesse. Questa è generalmente un'assunzione molto forte e non controllabile.
4. Il campione è abbastanza grande. Con pochi casi disponibili, il valutatore potrebbe dover accettare anche abbinamenti meno precisi, con un conseguente aumento della distorsione. In queste circostanze, gli effetti stimati potrebbero risentire della scelta della tecnica di abbinamento adottata.
5. L'assegnazione casuale non è un'opzione. Ci sono circostanze nelle quali l'assegnazione casuale è problematica e l'abbinamento offre alcuni vantaggi. Tuttavia, un'assegnazione casuale condotta propriamente elimina qualsiasi preoccupazione di distorsione dei risultati da selezione.

Ulteriori informazioni

EC – European Commission, Joint Research Center (2012), *The European Commission Joint Research Centre manual*. (http://bookshop.europa.eu).

Glennerster R., Takavarasha K. (2013), *Running randomized evaluations: A practical guide*. Princeton, NJ: Princeton University Press. pp. 8-12 e pp.73-77.

Khandker S.R., Koolwal G.B., Samad H.A. (2010), *The World Bank Handbook - Quantitative Methods and Practices*. Washington: The World Bank. (https://openknowledge.worldbank.org).

FASE 5

FASE 5 - ANALIZZARE E INTERPRETARE L'EFFETTO DELL'INTERVENTO

I vari metodi di valutazione dell'impatto stimano gli effetti di un intervento confrontando i risultati del gruppo che ha beneficiato dell'intervento stesso e del gruppo di controllo. L'effetto netto di un intervento corrisponde generalmente alla differenza nel livello della variabile d'interesse tra i due gruppi. Questo capitolo presenta una panoramica generale di alcuni aspetti importanti da considerare nell'interpretazione dei risultati, rilevanti per tutti i diversi metodi di valutazione.

La scelta del momento di misurazione dei risultati

Alcuni risultati possono richiedere un po' di tempo prima di manifestarsi e diventare pienamente osservabili dai ricercatori. Questo è il tipico caso di misure di attivazione, in cui le persone in cerca di lavoro sono incoraggiate a fermare temporaneamente la loro ricerca di lavoro per seguire la formazione che viene offerta loro dall'intervento (periodo di *lock-in*). La ricerca del lavoro inizia infatti solo dopo il completamento del ciclo di formazione, un processo che può a sua volta richiedere qualche tempo ulteriore prima che si producano risultati misurabili. Se i risultati (i tassi di occupazione) fossero misurati durante il periodo di *lock-in*, l'impatto sarebbe ovviamente sottovalutato. Per questo motivo è di vitale importanza scegliere adeguatamente il momento in cui misurare i risultati. Allo stesso modo, e più in generale, le politiche di investimento sociale possono richiedere uno sforzo a breve termine per generare benefici a più lungo termine. Mentre nel breve periodo il risultato di questa politica potrebbe non essere percepibile, esso diventa più evidente con il trascorrere del tempo.

Monitoraggio della conformità del programma (compliance), limitazione dell'attrito e garanzia di oggettività

I risultati possono essere fuorvianti se alcune delle unità assegnate al gruppo di controllo dovessero comunque ricevere il servizio previsto dal programma e/o alcuni beneficiari rimanerne esclusi. Una conformità parziale può potenzialmente ridurre la differenza tra i gruppi in termini di esposizione all'intervento. Un caso estremo potrebbe verificarsi nel caso in cui il medesimo numero di soggetti nei due gruppi ricevesse il programma. In questo scenario, sarebbe impossibile valutare l'impatto

dell'intervento, perché entrambi i gruppi avrebbero avuto la stessa esposizione al programma. Solo con il monitoraggio continuo della conformità del programma mentre l'intervento è in corso di attuazione, i ricercatori possono agire tempestivamente per garantire il rispetto e la qualità dei risultati. Oltre a questo, i tassi di conformità devono essere rigorosamente registrati in modo che possano essere presi in considerazione nell'analisi dei risultati. Un altro aspetto importante da considerare quando si interpretano i risultati è l'attrito del programma. L'attrito è ciò che si verifica quando i ricercatori non sono in grado di misurare i risultati per alcuni dei soggetti inclusi nella valutazione. Se tipo e dimensione dell'attrito sono diversi nel gruppo dei beneficiari e in quello di controllo, i risultati potrebbero essere distorti. Si consideri, per esempio, una politica di attivazione di successo, nella quale un gruppo di persone in cerca di lavoro riceva un intenso programma di consulenza per la ricerca di occupazione (gruppo dei beneficiari) e un altro gruppo di senza lavoro possa accedere solo al programma di consulenza standard (gruppo di controllo). I disoccupati meno impiegabili nel gruppo di intervento potrebbero essere più propensi a migliorare le loro prospettive di lavoro e a non abbandonare il programma, mentre le persone meno impiegabili nel gruppo di controllo potrebbero essere scoraggiate e abbandonare l'intervento. La conseguenza potrebbe essere una sovra-rappresentazione delle persone meno occupabili nel gruppo di intervento che renderebbe i due gruppi non più confrontabili. In questo caso, l'impatto del programma risulterebbe sottovalutato. Infine, è della massima importanza che la valutazione sia condotta da una parte terza obiettiva e indipendente. Il disegno valutativo e i suoi metodi devono essere accuratamente registrati e, quando possibile, i dati devono essere resi pubblici per facilitare la replicabilità della valutazione.

Comunicazione dei risultati

Valutazioni indipendenti e imparziali, con sufficiente potenza (statistica) e con progettazione adeguata possono anche registrare effetti nulli. Un impatto nullo può essere utile quanto un ampio impatto positivo o negativo. Per questo motivo, i ricercatori e i *policy makers* dovrebbero avere ben chiaro che l'obiettivo di una valutazione di impatto è misurare se un intervento sia efficace nel raggiungere gli obiettivi previsti e che è comunque possibile che l'intervento non mostri alcun effetto o addirittura ne mostri di negativi. Al fine di migliorare l'apprendimento reciproco tra i membri della comunità di riferimento, il gruppo di valutazione deve riportare e diffondere i risultati in modo trasparente e completo.

Ulteriori informazioni

Gertler P.J., Martinez S., Premand P., Rawlings L.B., Vermeersch C.M. (2010), *Impact evaluation in practice*. Washington, DC: World Bank. (http://web.worldbank.org)

World Bank, World Bank Impact Evaluation Toolkit, Module 7: Analyzing Data and Disseminating Results. (http://web.worldbank.org).

FASE 6

FASE 6 - DISSEMINARE I RISULTATI

Quando i risultati della valutazione hanno importanti implicazioni sul piano delle politiche, la ricerca dovrebbe alimentare il processo di costruzione delle politiche. Offrendo ai responsabili politici la possibilità di costruire interventi basati sui risultati misurati, siano essi negativi o positivi, si offre loro la possibilità di migliorare ulteriormente il loro impatto. Questo capitolo fornisce alcuni consigli su come diffondere i risultati della valutazione.

Capire la rilevanza politica di una valutazione

La rilevanza di una politica è molto dipendente dal tempo: un argomento potrebbe essere caldo un giorno e freddo la settimana successiva. Pertanto, è importante tenere d'occhio l'agenda politica. La finestra di opportunità può nascere, per esempio, nel corso delle discussioni di bilancio, quando i politici impostano le loro priorità e allocano le risorse. Una prova fornita al momento giusto sarà più probabilmente presa in considerazione e trasformata in politica.

Diffondere i risultati in un formato accessibile

Al di là dei risultati della ricerca, è anche importante comunicare le implicazioni della valutazione / sperimentazione della politica. I risultati delle valutazioni di impatto sono spesso presentate *in working paper* o riviste accademiche, documenti che tendono ad essere scritti in modo molto tecnico che può limitarne la circolazione tra i responsabili politici. Una responsabilità chiave è dunque quella di rendere la ricerca più accessibile, estraendo i risultati più interessanti da documenti più lunghi per condensarli in relazioni e presentazioni scritte in linguaggio non tecnico.

Diffondere anche i dettagli

Mentre la maggior parte degli utenti finali sarà raggiunta con dei comunicati sintetici, rimane comunque utile la diffusione di una documentazione completa del progetto, compresi i dati che sono stati raccolti, eventualmente dopo essere stati resi anonimi. Altri ricercatori potrebbero infatti essere interessati al lavoro fatto e la loro attenta

lettura aumenterà l'affidabilità e l'estensione dei risultati. Anche quando i risultati sono stati doverosamente verificati attraverso i meccanismi di peer-review, permettere ad altri di setacciare i dati può produrre nuove intuizioni. Infine, la piena diffusione dei rapporti di ricerca aiuta a produrre rassegne sistematiche e meta-valutazioni.

Pubblicare i risultati nei registri di valutazione

I responsabili delle politiche sociali e gli operatori possono a volte avere difficoltà a trovare l'evidenza prodotta, che rimane spesso confinata in riviste accademiche. Un certo numero di organizzazioni hanno compiuto degli sforzi per rendere disponibile in un unico posto l'evidenza prodotta rigorosamente. A questo proposito, si segnalano alcuni luoghi deputati alla raccolta di tali materiali:

- piattaforma europea finanziata dall'UE relativa all'infanzia (*European Platform for Investing in Children*)[9];
- *J-PAL Evaluation Database*[10];
- *Evaluation Database of the Coalition for Evidence-Based Policy*[11];
- Evaluation Database of the Network of Networks in Impact Evaluation (NONIE)[12].

Ulteriori informazioni

DFID Research Uptake Guidance. (www.gov.uk).

Iqbal D., Tulloch C., From Research to Policy: Using Evidence from Impact Evaluations to Inform Development Policy. Cambridge, MA: MIT, J-PAL, Department of Economics. (www.povertyactionlab.org).

Policy Impact Toolkit. (http://policyimpacttoolkit.squarespace.com).

Stachowiak S. (2009), Pathways for Change: 6 Theories about How Policy Change Happens. Seattle, WA: ORS, Organizational Research Services. (http://goo.gl)

9 http://europa.eu/epic/index_en.htm

10 http://www.povertyactionlab.org/search/apachesolr_search?filters=type:evaluation

11 http://toptierevidence.org/.

12 http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf.

FASE 7

FASE 7 - DAL LOCALE AL GLOBALE

Come si fa a sapere se un programma che ha mostrato di essere efficace in una fase pilota ha lo stesso impatto se esteso o riprodotto in un luogo differente? Si tratta di una questione di primaria importanza, definita validità esterna. La validità esterna, nota anche come 'generalizzabilità', è il grado di certezza che i risultati ottenuti in un contesto specifico lo siano anche in contesti differenti. Questo capitolo spiega come interventi efficaci possano essere estesi in modo che il nuovo approccio generi un impatto reale e diventi prassi[13].

La sfida della trasferibilità dei risultati

La validità interna della valutazione è un presupposto importante (anche se non sufficiente) per la generalizzabilità dei risultati. Se non è possibile essere certi che la valutazione misuri il vero impatto del programma in un contesto specifico, allora sarà assai più difficile generalizzare le conclusioni a un contesto diverso. Uno dei vantaggi delle valutazioni randomizzate è proprio la loro forte validità interna. Come si è visto, l'assegnazione casuale garantisce che l'unica differenza tra i gruppi di intervento e di controllo consista nel ricevere o meno l'intervento. Eventuali modifiche nei risultati possono quindi essere tranquillamente attribuite all'intervento in fase di valutazione, senza la necessità di fare inferenze aggiuntive sulla comparabilità dei gruppi.

Ci sono quattro fattori principali che influenzano la generalizzabilità di una valutazione, in particolare la qualità della realizzazione, la sua scala, il contesto e il contenuto specifico del programma.

1. La qualità della realizzazione: i programmi pilota sono spesso attuati con grande cura e con personale ben addestrato. Può essere difficile mantenere gli stessi standard su una scala più ampia. I ricercatori dovrebbero attuare gli interventi in luoghi rappresentativi, con partner rappresentativi e campioni rappresentativi dei beneficiari reali.

13 EC - European Commission (2013), *Guide to social innovation*. DG Regional and Urban Policy and DG Employment. Brussels: European Commission.

2. La scala di realizzazione: un programma attuato su piccola scala può avere effetti diversi una volta ampliato (effetti di equilibrio generale). I ricercatori possono adattare il disegno della valutazione in modo da catturare questi effetti utilizzando un numero di unità di osservazione sufficientemente ampio (ad esempio a livello di comunità)[14]. Confrontare i risultati nelle comunità che hanno introdotto il programma con quelli delle comunità che non lo hanno fatto può aiutare a identificare e misurare alcuni di questi effetti.
3. Il contesto di implementazione: Un intervento che dimostra di essere efficace in un contesto può avere un impatto diverso in un altro contesto istituzionale e culturale. La teoria comportamentale può aiutare a definire quali aspetti del contesto rischiano di essere rilevanti per un particolare programma.
4. Il contenuto del programma: gli effetti di un determinato programma possono variare se alcuni dei suoi componenti vengono modificati.

Per saperne di più

Hunt A., Mullainathan S. (2012), *External validity and partner selection bias*. Cambridge, MA: National Bureau of Economic Research, NBER Working Paper n. 18373.

Iqbal D., Duflo E., Glennerster R., Tulloch C. (2012), *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education*. Cambridge, MA: MIT, J-PAL, Department of Economics. (www.povertyactionlab.org).

Larry C., Kohl R. (2006), *Scaling Up-From Vision to Large-scale Change, A Management Framework for Practitioners*. (www.msiworldwide.com).

14 Alcuni affiliati a J-PAL hanno intrapreso una valutazione sugli effetti di un programma di collocamento e di consulenza al lavoro e di spostamento intensivi. Per ulteriori informazioni: http://www.povertyactionlab.org/publication/job-placement-and-displacement

II. PARTE - CASI STUDIO

Questa parte illustrerà il ruolo della metodologia nella valutazione dei risultati di alcuni esempi concreti di riforme sistemiche. I passaggi logici presentati in precedenza sono quindi messi in pratica in casi concreti al fine di aiutare il lettore a comprendere meglio come le politiche sociali innovative possano essere sostenute attraverso le evidenze della ricerca.

I casi studio che seguono hanno lo scopo di ancorare la metodologia a cambiamenti concreti e realistici della politica sociale. Questi cambiamenti sono plausibili e persino auspicabili ma, mentre scopo e principi sono ben definiti, l'effettivo raggiungimento del loro obiettivo potenziale potrebbe dipendere da dettagli nell'attuazione degli interventi. Le riforme delle politiche che saranno illustrate di seguito insieme ad alcune affidabili misure dei risultati raggiunti, contribuiranno a massimizzarne l'utilità. Gli esempi che seguono mostrano il valore aggiunto di accompagnare le riforme con opportune valutazioni degli effetti.

ESEMPIO 1

ESEMPIO 1 - COME VALUTARE UNA RIFORMA DEGLI ASSEGNI DI INVALIDITÀ

1. Introduzione

Questa nota illustra come valutare l'impatto di una riforma dell'assicurazione contro l'invalidità. Si presentano i principali disegni di ricerca disponibili allo scopo e l'uso che se ne è fatto in passato. Questi disegni variano essenzialmente per:

1. i requisiti metodologici; e
2. le ipotesi che devono essere fatte per quanto riguarda la comparabilità dei gruppi di intervento e di confronto.

Questa discussione è illustrata con esempi tratti da vari Paesi - Estonia, Norvegia, Regno Unito e Danimarca - che hanno progettato o realizzato riforme per stimolare l'attivazione dei beneficiari degli assegni per l'inabilità al lavoro e che hanno accompagnato tali riforme con rigorosi piani di valutazione. La nota presenta i vincoli che hanno determinato le scelte di specifici disegni e metodologie di valutazione ed è organizzata come segue: il paragrafo 2 descrive brevemente le caratteristiche della riforma, il paragrafo 3 spiega come costruire situazioni controfattuali, il paragrafo 4 discute i diversi metodi che potrebbero essere applicati e li illustra con esempi concreti e, infine, il paragrafo 5 fornisce un esempio dei vincoli contestuali che possono sorgere durante il processo di valutazione.

2. Uno sguardo alle riforme delle assicurazioni contro l'invalidità

Negli ultimi decenni, il numero di beneficiari di assegni di invalidità è aumentato, in particolare nei Paesi del Nord Europa (OECD, 2009). Questa progressione potrebbe essere il risultato di un deterioramento delle opportunità offerte dal mercato del lavoro insieme a politiche sociali che combinano prestazioni di invalidità generose con azioni di screening e monitoraggio indulgenti. In realtà questo tipo di prestazione sociale è molto più generoso dei sussidi contro la disoccupazione di lungo periodo

e risulta quindi comparativamente più attraente[15]. Il costante aumento del numero di beneficiari di questa misura può ostacolare la sostenibilità del sistema pensionistico e generare fenomeni di carenza di manodopera. Le misure finalizzate a ridurre ingressi e dimensioni dello stock dei beneficiari di assegni di invalidità comprendono:

- il rafforzamento dello *screening* per l'acceso alle prestazioni
- la riduzione del livello delle prestazioni, e
- l'aumento delle uscite dal regime, sia attraverso incentivi fiscali, sia con altre forme di attivazione.

La riforma del modello contempla tutte queste opzioni con una varietà di possibili disposizioni di attuazione.

3. Criteri per valutare la riforma

L'obiettivo della riforma dovrebbe essere duplice: frenare l'aumento del numero di beneficiari e rendere attivi i soggetti solo parzialmente disabili. La quota delle persone disabili inserite nel mondo del lavoro e il numero dei soggetti che passa dall'assegno di invalidità a un posto di lavoro sono due importanti indicatori del successo delle politiche di attivazione. La partecipazione delle imprese è un fattore determinante di questo successo. A questo proposito, è interessante valutare la "scrematura" che si verifica quando le imprese selezionano i soggetti più occupabili tra i disabili e lasciano invece i meno impiegabili ai lavori socialmente utili o a ad altri posti di lavoro sussidiati con risorse pubbliche. Ciò richiede che si misuri il numero di persone disabili impiegate nel settore privato (invece che nella pubblica amministrazione) e la quota di posti di lavoro sussidiati.

In generale, i principali **risultati** su cui valutare tale riforma sono:

- le dimensioni e la composizione degli ingressi nel sistema dei sussidi;
- le ore di lavoro, il reddito, la ripartizione tra settore pubblico e privato (e/o altre misure di qualità dell'occupazione) dei soggetti inseriti nel sistema dei sussidi;
- il numero e la destinazione delle uscite dal sistema dei sussidi (verso il lavoro, verso lavori più impegnativi);
- il valore dei sussidi, il reddito e il rischio individuale di povertà;
- la salute dei partecipanti;
- il costo della politica.

La raccolta dei dati, da fonti amministrative o da indagini *ad-hoc*, deve essere pianificata in anticipo, prima dell'applicazione delle nuove misure, e adattata al protocollo di valutazione.

Una domanda di valutazione generale potrebbe essere: come funziona il sistema? considerando tutte le sue diverse caratteristiche, fa la differenza? Altre questioni importanti sono: quale misura della politica risulta più efficace? Tutte le misure sono necessarie?

15 Questa tendenza è stata documentata negli Stati Uniti anche da Autor e Duggan, 2003.

I **meccanismi** della riforma dipendono da vari aspetti, i più importanti dei quali sono:

- Quanto diverso sarà il nuovo processo di *screening*?
- Quanto i destinatari della riforma sono sensibili al livello dei sussidi rispetto ai potenziali salari guadagnati sul mercato del lavoro o ad altri benefici? Più precisamente, quanto è importante il profilo degli incentivi per la ri-occupazione?
- Quanto sono efficienti le politiche di attivazione? Per quanto tempo sono offerte? Qual è il tasso di adesione a queste misure?

Nessuno dei programmi di valutazione che analizzeremo affronta o tiene conto contemporaneamente di tutte queste questioni o considera contemporaneamente tutti i risultati possibili. Alcuni protocolli di valutazione forniscono una stima dell'impatto del sistema come insieme di diverse misure. Altri protocolli forniscono invece una stima degli effetti delle singole misure e, in questo modo, contribuiscono a individuare le opzioni più efficaci. Allo stesso modo, alcuni protocolli aiutano a stimare l'impatto sui flussi di accesso al sistema, mentre altri sono più adatti a misurare gli effetti individuali. L'analisi che segue fa riferimento a valutazioni che mettono a confronto il nuovo regime con quello esistente.

4. Come costruire controfattuali

Per valutare se la riforma fa la differenza, di quanto e con quali costi e benefici, si ha la necessità di stabilire una situazione controfattuale[16], una stima del risultato (o risultati) in assenza della nuova politica.

Supponiamo che, senza la riforma, ci si attenda che il numero dei beneficiari degli assegni di invalidità aumenti e che sia disponibile una stima di tale aumento nei prossimi dieci anni.

Non è però possibile valutare l'impatto della riforma semplicemente misurando il numero di beneficiari degli assegni prima e dopo l'attuazione della riforma, perché queste cifre possono cambiare nel corso del tempo per un gran numero di motivi indipendenti dalla riforma stessa. Ad esempio, una diminuzione del tasso di disoccupazione può influenzare (riducendole) le richieste di assegni di invalidità, un risultato che sarebbe sbagliato attribuire alla riforma. Al contrario, inattesi shock economici negativi possono aumentare il numero dei richiedenti, anche se la riforma dimostra di essere altamente efficace. In breve, raggiungere o meno un determinato numero di beneficiari di assegni di invalidità dice poco sulla desiderabilità della riforma: la variazione del numero di beneficiari deve essere infatti confrontata con una situazione controfattuale ben definita. La sfida è dunque quella di individuare o costruire tale controfattuale. Ci sono due possibili livelli di analisi, il controfattuale a livello individuale e quello a livello di popolazione. Ciascuno di questi livelli consente la misurazione dell'impatto su risultati differenti.

16 Si veda a proposito; ESF Guide on Counterfactual Evaluation: Design and Commissioning of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2012.

4.1. Controfattuali a livello individuale

L'introduzione di un nuovo regime modifica le condizioni con cui le persone si confrontano. Le modifiche possono riguardare nuove tariffe, misure di attivazione e incentivi per l'occupazione. Queste novità potrebbero avere un impatto sulla maggior parte dei risultati mostrati in precedenza, come i tassi di partecipazione al mercato del lavoro, quelli di uscita dal regime degli assegni di invalidità, il livello di reddito e la salute dei beneficiari. Per stimare l'impatto della riforma su ognuno di questi risultati, si ha la necessità di confrontare i beneficiari delle nuove misure con beneficiari con caratteristiche simili sottoposti alle condizioni in vigore prima della riforma. Controfattuali a livello individuale possono essere utilizzati anche per stimare l'impatto dei diversi aspetti della riforma e, in questo modo, per individuare quelli più efficienti. Ciò è molto importante dal punto di vista della politica, in quanto la spesa pubblica può essere ottimizzata puntando sulle opzioni che dimostrano di avere effetto sui risultati desiderati. Inoltre, controfattuali a livello individuale aiutano a stimare l'eterogeneità dell'impatto su diverse sotto-popolazioni favorendo un eventuale segmentazione dell'intervento. Ad esempio, affrontare la questione della "scrematura" da parte dei datori di lavoro richiede una focalizzazione sull'occupabilità degli individui; il loro reinserimento nel mercato del lavoro potrebbe risultare differente nelle condizioni pre e post riforma.

4.2. Controfattuali a livello di popolazione

La riforma può anche influenzare il numero e le caratteristiche delle persone in ingresso, attraverso due meccanismi contemporanei: il cambiamento del processo di *screening* (sul lato dell'offerta) e la variazione del valore che i potenziali candidati attribuiscono al sussidio (sul lato della domanda). Per misurare la variazione del flusso, il controfattuale non può essere basato sul confronto di individui simili nel nuovo e nel vecchio schema, ma deve essere calcolato a livello della popolazione potenzialmente ammissibile a ciascun regime, perché il flusso in ingresso è misurato come quota di una particolare popolazione. Occorre quindi osservare due popolazioni simili in contesti simili e l'unica differenza tra le due popolazioni deve essere il regime (vecchio o nuovo) cui possono accedere[17]. Dato che l'obiettivo è misurare come il nuovo regime incide sull'ottenimento del sussidio , le popolazioni da confrontare non sono composte da beneficiari, ma da un insieme ben definito di potenziali candidati che potrebbero entrare nel sistema e che, a seconda delle regole, sceglieranno se farlo o meno.

Utilizzando controfattuali a livello di popolazione, è possibile misurare i nuovi tassi di partecipazione al mercato del lavoro, il numero di persone che cessano di accedere alle prestazioni (tasso di uscita) e il livello di reddito dei beneficiari che entrano nel nuovo regime e confrontarli con quelli dei beneficiari che entrano nel programma iniziale. Tuttavia, queste misure si combinano con effetti di composizione (persone diverse possono reagire in modo diverso a un particolare regime) ed effetti di regime (le stesse persone si comportano in modo diverso in base alle nuove regole). In altre parole, se il nuovo regime incide sull'afflusso dei beneficiari, gli individui che parteci-

17 La differenza tra controfattuali a livello di popolazione e a livello individuale è che nel primo caso si devono trovare due popolazioni simili di potenziali beneficiari, la prima sottoposta al regime iniziale e l'altra al nuovo regime; nel secondo caso, si devono invece trovare singoli beneficiari simili, alcuni che partecipano al programma iniziale e altri al nuovo schema.



pano al nuovo schema non sono più paragonabili a quelli del regime iniziale, sia perché hanno caratteristiche diverse, sia perché si trovano ad affrontare regole diverse.

Al fine di effettuare un valido confronto tra individui nel nuovo e nel vecchio regime, sarebbe necessario valutare quanto accade con i beneficiari del regime iniziale esclusi dal nuovo regime e il gruppo di quelli che sono ammissibili solo al nuovo schema[18]. Tuttavia, non è materialmente possibile identificare i soggetti che sarebbero comunque entrati in entrambi i sistemi e quelli che sarebbero entrati nel regime iniziale ma non in quello nuovo, perché molte delle caratteristiche che determinano l'entrata non sono osservabili. Le caratteristiche non osservabili possono includere la motivazione al rientro al lavoro e la sensibilità al livello degli assegni.

In breve, controfattuali individuali e di popolazione misurano risultati differenti. I controfattuali a livello individuale sono più adatti per valutare l'impatto sui risultati, come i nuovi tassi di partecipazione al mercato del lavoro, i tassi di uscita, il livello di reddito e salute dei beneficiari, mentre i controfattuali a livello di popolazione consentono di determinare l'impatto della riforma sulla dimensione e la composizione degli ingressi e dello *stock* dei beneficiari. È particolarmente difficile valutare tutte le implicazioni della riforma considerate, perché vi sono molti elementi che influenzano sia l'ingresso sia la condizione dei beneficiari.

18 Un'analisi completa degli effetti generati dalla riforma sarebbe un'impresa formidabile. Sarebbe, infatti, necessario seguire una popolazione molto ampia (tutte le persone con una probabilità maggiore di zero di beneficiare degli assegni). Ciò è fattibile, in linea di massima, applicando il metodo della Differenza nelle differenze (DID) o approcci simili, ma l'esercizio non sarà svolto in questa sede. In questo caso consideriamo alcuni importanti insiemi di risultati solo separatamente.

Si deve notare che, indipendentemente dal livello di analisi, una riforma di questo tipo deve essere accompagnata sia da un *follow-up* amministrativo, che registra le persone dentro e fuori del sistema, sia da indagini campionarie che studiano più da vicino piccoli gruppi, raccogliendo informazioni su elementi non registrati dalle fonti amministrative.

5. Potenziale dei diversi metodi di valutazione d'impatto controfattuale

5.1. Abbinamento statistico

Descrizione

Questo metodo costruisce un gruppo di confronto facendo corrispondere ciascuno dei beneficiari del nuovo regime a un beneficiario simile del vecchio regime, utilizzando un insieme di caratteristiche osservabili. Un abbinamento di successo richiede una ricerca preliminare per identificare le variabili che potrebbero essere statisticamente correlate alla probabilità di partecipare al programma e ai risultati d'interesse. Per creare un numero sufficiente di corrispondenze sono quindi necessari campioni molto grandi. Questo metodo fornisce una stima dell'effetto di un intervento per tutti i nuovi partecipanti al nuovo schema che possono essere abbinati a un beneficiario che gode delle condizioni precedenti.

L'accoppiamento potrebbe essere svolto per confrontare i beneficiari del nuovo e del vecchio sistema. Si noti, tuttavia, che questo metodo ha un interesse molto limitato se il regime iniziale viene completamente sostituito da quello nuovo e quindi il confronto può basarsi solo su dati retrospettivi. In questo scenario, i partecipanti al nuovo schema sono inevitabilmente diversi da quelli nel regime iniziale, almeno in un aspetto importante: il contesto economico che si trovano ad affrontare.

Assunzioni

L'ipotesi principale è che tutte le caratteristiche di base che influenzano la partecipazione e i risultati di interesse possano essere osservate e valutate.

Esempio

> Nel contesto della valutazione del progetto "Pathways to work" gestito dall'*Institute for Fiscal Studies* (vedi oltre), Adam, Bozio e Emmerson (2009) hanno discusso dell'adozione di questo metodo per valutare il pacchetto *Choices*, uno dei componenti del programma. *Choices* include una varietà di sistemi volontari volti a migliorare l'occupabilità e le prospettive di lavoro dei richiedenti. I ricercatori hanno concluso che l'abbinamento, in base a un gran numero di caratteristiche osservabili di partecipanti a non partecipanti al programma, non fosse una strategia di valutazione rigorosa. Secondo gli studiosi sarebbero infatti rimaste escluse troppe importanti caratteristiche non osservabili, il che avrebbe reso impossibile sapere quanta parte delle scelte degli individui avrebbe potuto essere determinata dalle loro caratteristiche non osservabili e quanta dalle specifiche del programma *Choices*.

Applicabilità alla riforma ipotizzata

Se applicato alla riforma di cui stiamo trattando, questo metodo equivarrebbe a confrontare individui con caratteristiche osservabili simili o molto vicine inseriti nelle due versioni dello schema. Il metodo consentirebbe di stimare l'impatto delle nuove caratteristiche dello schema sulle scelte individuali di adesione al nuovo regime. Per stimare l'effetto sul flusso, sarebbe necessario abbinare persone simili che si sono trovate di fronte alle due versioni differenti dello schema e confrontare le probabilità di ingresso e di uscita dal programma.

In teoria, l'abbinamento potrebbe essere effettuato per valutare la riforma su scala completa, a condizione che siano disponibili i dati (di fonte amministrativa o di una rilevazione organizzata *ad-hoc*) su una coorte precedente e una coorte successiva alla data di attuazione. In questo caso, le informazioni richieste dovrebbero includere il grado e il tipo di disabilità, l'età, il sesso, lo *status* socio economico familiare, il curriculum accademico e lavorativo. La causa della disabilità (ad esempio, il lavoro) potrebbe essere molto importante. Queste informazioni dovrebbero essere raccolte per ogni individuo, inclusi quelli che non hanno presentato domanda di adesione.

Tuttavia, come già ricordato, questo metodo richiede che siano osservabili e rappresentabili tutte le caratteristiche che possono influenzare le scelte di partecipazione e i risultati di interesse. In generale, si tratta di un'ipotesi forte e, in questo caso particolare, altamente improbabile. Infatti, molte delle caratteristiche rilevanti che potrebbero influenzare i risultati finali sono legate alla capacità lavorativa e al livello di disabilità. Anche se i dati amministrativi fornissero una buona *proxy* della capacità lavorativa, uno degli elementi centrali della riforma riguarda però proprio il cambiamento del processo di selezione. Questo implica che le due misure di disabilità (valutazione della capacità lavorativa nel vecchio e nel nuovo sistema di selezione) sarebbero difficili da confrontare. Se invece la riforma non avesse modificato il processo di selezione, l'approccio potrebbe essere più interessante.

Infine, abbinando soggetti sulla base di variabili misurate in modo uniforme nel vecchio e nel nuovo schema (età, sesso, istruzione, etc.) risulta possibile scomporre i cambiamenti complessivi osservati sui beneficiari (tasso di occupazione, etc.) distinguendo tra l'effetto di composizione (attribuibile alle variabili utilizzate nell'abbinamento) e gli altri effetti *complessivi*, compreso l'impatto della riforma, ma anche delle caratteristiche che non sono state prese in considerazione nell'abbinamento[19].

L'abbinamento statistico potrebbe consentire anche di confrontare l'effetto di varie misure di attivazione (o la loro mancanza) sui soli individui che partecipano al nuovo schema, al posto di confrontare vecchio e nuovo schema.

5.2. Confronto attorno al punto di discontinuità (RDD)

Descrizione

Questo metodo confronta individui posizionati intorno (appena al di sopra o appena al di sotto) di una determinata soglia di ammissibilità misurata da una variabile con-

19 Nella letteratura statistica questo metodo è chiamato decomposizione "Oaxaca-Blinder".

tinua. Infatti, tali individui sono probabilmente molto simili e la soglia determina la loro esposizione o meno all'intervento. La larghezza di banda tra limite inferiore e superiore (all'interno dei quali è posizionata la soglia) determina la dimensione del campione.

Assunzioni

Questo metodo si basa sul presupposto che l'ammissibilità all'intervento si fondi su un criterio di selezione chiaramente quantificabile basato su un punteggio continuo e che i partecipanti non siano in grado di anticipare e manipolare i punteggi vicino al punto di soglia. Assume inoltre che gli individui appena sotto e appena sopra la soglia non siano significativamente diversi.

Esempio

Kostol e Mogstad (2014) hanno usato questo metodo per valutare l'impatto di un cambiamento degli incentivi al lavoro ai beneficiari di assegni di invalidità in Norvegia. Gli individui che erano stati assegnatari di assegni prima del 1 gennaio 2004 erano stati esposti a norme più generose riguardo alla possibilità di percepire congiuntamente assegni di invalidità e salari rispetto alle persone entrate successivamente a tale data. Gli autori hanno ipotizzato che le persone entrate subito prima e subito dopo quella data fossero molto simili tra loro e, quindi, che le differenze nei risultati (ad esempio il lavoro durante il periodo di invalidità, i tassi di uscita, etc.) avrebbero potuto essere attribuite alla variazione delle norme sugli incentivi al lavoro. Non si tratta di un confronto prima e dopo, perché tutti gli individui sono osservati simultaneamente e, quindi, nello stesso contesto economico generale. Gli autori hanno scoperto che gli incentivi finanziari inducono una parte sostanziale dei beneficiari degli assegni a rientrare al lavoro, ma solo quelli più giovani. Questo conferma l'idea che alcuni beneficiari di assegni di invalidità possano effettivamente lavorare e che gli incentivi sono efficaci nell'incoraggiarli a farlo.

Un aspetto essenziale di questo metodo, fortemente sottolineato da Kostol e Mogstad (2014) è che i beneficiari hanno avuto accesso al programma prima del cambiamento delle regole loro applicate. Gli individui entrati a cavallo del 1 gennaio 2004 non erano quindi consapevoli che ci sarebbe stato un cambiamento delle regole e sono stati ammessi con lo stesso meccanismo di selezione. Trattandosi di un cambiamento retroattivo, le persone non erano in grado di modificare le proprie scelte di ingresso nel programma.

Nel caso in cui il processo di selezione e/o il valore dell'assegno fossero stati modificati insieme al contenuto del programma, l'ipotesi di utilizzare il confronto attorno alla soglia (RDD) non avrebbe invece più retto: le persone appena prima e appena dopo la data sarebbero state molto diverse fra loro e il confronto sarebbe quindi stato privo di significato.

Applicabilità alla riforma ipotizzata

Il requisito principale per utilizzare questo metodo nella valutazione della riforma è che ci sia qualche variabile continua che determina l'ingresso nel vecchio o nel nuovo regime. Come evidenziato da Kostol e Mogstad (2014), un candidato naturale a questo ruolo è la data di applicazione della riforma, a condizione che si considerino gli individui entrati in prossimità di tale data. Il metodo potrebbe essere utilizzato per stimare congiuntamente l'impatto di incentivi finanziari e di meccanismi di attivazione, ma non sarà comunque in grado di distinguere l'impatto dei singoli componenti. Inoltre,

devono essere mantenute due condizioni: gli incentivi finanziari e le misure di attivazione devono essere introdotte separatamente dal cambiamento nel processo di selezione e la riforma deve essere introdotta retroattivamente su determinato gruppo di beneficiari. Non è chiaro che questo sarebbe applicabile alla riforma.

Se il nuovo regime è basato su soglie, offre cioè benefici a seconda del grado di disabilità, misurata e riportata su un continuum, l'impatto del programma sui singoli individui potrebbe essere valutato confrontando i comportamenti e gli esiti lavorativi dei soggetti posizionati appena al di sotto e appena al di sopra della soglia (o delle soglie). Si noti tuttavia, che ciò non significa confrontare il vecchio e il nuovo schema, ma l'impatto di essere o meno inseriti nello schema.

Alcuni dei limiti di questo metodo consistono proprio nell'impossibilità di stimare impatti sugli ingressi. Inoltre, la sua applicazione richiede un flusso in ingresso consistente, in modo che siano disponibili almeno alcune centinaia di individui vicini alla data o al livello di disabilità scelto come punto di discontinuità.

5.3. Differenza nelle differenze (DID)

Descrizione

Il metodo differenza nelle differenze (DID) richiede l'introduzione di una dimensione sperimentale nel programma sotto forma di aree pilota e di controllo. Questo metodo confronta la variazione dei risultati prima e dopo l'inizio del programma, nelle aree pilota e in quelle di controllo. In questo modo si fornisce una misura dell'impatto per l'intera popolazione dei partecipanti controllando per le condizioni costanti (osservate e non) che possono essere correlate sia con i risultati finali sia con l'appartenenza al gruppo di controllo.

Assunzioni

L'approccio è basato sull'assunzione delle "dinamiche parallele" nel tempo. Al fine di determinare se la differenza nei risultati è dovuta al programma, si deve infatti assumere che l'evoluzione delle variabili risultato dei partecipanti e dei non partecipanti sarebbero identiche in assenza del programma. Un modo per validare l'ipotesi è verificare se entrambi i gruppi hanno mostrato andamenti paralleli nelle variabili risultato prima dell'introduzione del programma. Altre modalità di verifica prevedono, da un lato, l'esecuzione di *test* che simulano trattamenti *placebo* su finti gruppi di trattamento (nessuno dei quali è influenzato dall'avvio della riforma) oppure stimano gli effetti su risultati che non dovrebbero essere ragionevolmente attribuibili all'intervento.

Esempio

La valutazione del programma britannico "Pathways to Work" ha utilizzato questo approccio per verificare l'impatto del programma sui flussi di passaggio dagli assegni di invalidità all'occupazione (Adam, Bozio e Emmerson, 2010). Questa riforma sperimentale comprendeva forti incentivi economici per il rientro al lavoro, il monitoraggio (interviste obbligatorie) e l'attivazione (sistemi di consulenza volontari - tra cui il programma *Choices* di cui sopra). La riforma è stata sperimentalmente introdotta in maniera progressiva nei vari distretti. I funzionari del Dipartimento per il Lavoro e le Pensioni (DWP) hanno selezionato i distretti pilota prima

dell'intervento e hanno lasciato ai valutatori la possibilità di selezionare i distretti di controllo più adeguati sulla base di un insieme di caratteristiche osservabili a livello aggregato. Essendo una misura con variazioni attese a livello di quartiere, il controfattuale è stato costruito a livello di popolazione. La differenza di risultati riscontrati dopo l'attuazione della politica tra le aree pilota e quelle di controllo è stata interpretata come un effetto del programma *Pathways to Work*.

Da un punto di vista gestionale, non si può escludere che le aree pilota siano state selezionate perché "atipiche". Nella fattispecie, sono state scelte aree pilota nelle quali, per qualche tempo, era già stato adottato il programma Jobcentre Plus. A parità di altre condizioni, la selezione di un'area ad alte prestazioni come sito pilota aumenta la probabilità di osservare un risultato positivo. Sceglierne una meno performante, potrebbe essere visto come un modo di sfidare o cambiare una zona a basso rendimento. Entrambe le decisioni compromettono la comparabilità. Inoltre, selezionare le aree pilota sulla base della loro *performance* riduce la validità esterna della valutazione. Se la politica verrà implementata in aree con *performance* diverse, l'impatto del programma a regime sarà probabilmente diverso da quello del pilota.

L'ipotesi delle "dinamiche parallele" nel tempo diventa difficile da giustificare se le aree pilota e di controllo sono differenti (ad esempio perché i servizi sociali sono risultati migliori nelle aree pilota). Questa ipotesi non è verificabile, ma gli autori forniscono alcuni elementi interessanti. Hanno infatti scelto due gruppi di distretti pilota che hanno iniziato l'esperimento in due momenti separati. Gli autori hanno scoperto che prima dell'introduzione del programma, le aree di controllo e quelle pilota avevano tassi di uscita dal regime degli assegni di invalidità molto simili. I tassi di uscita si sono invece modificati solo dopo che le aree pilota sono entrate nel sistema e tale fenomeno è stato osservato separatamente durante la prima e la seconda ondata. Questa coincidenza conferma la bontà della metodologia scelta per la valutazione del programma e l'efficacia dello stesso.

I valutatori hanno rilevato che la riforma accelerava l'uscita dal sistema degli assegni d'invalidità, ma solo per coloro che ne sarebbero comunque usciti entro un anno. Tuttavia, hanno osservato anche effetti duraturi sull'occupazione. Questi due risultati sono stati interpretati attribuendo l'effetto alle scelte delle donne sposate che sarebbero comunque uscite dal sistema, potendo contare sulle risorse del *partner*, ma che sono invece tornate al lavoro proprio grazie al programma *Pathways*. Questa interpretazione era anche (pur debolmente) supportata dall'analisi dei dati per sottogruppi di popolazione.

Uno dei vantaggi del metodo della Differenza nelle differenze è che controlla i risultati per le differenze che non variano nel tempo, sia nelle caratteristiche osservabili sia in quelle non osservabili. Se i tassi d'ingresso nel nuovo regime rimangono invariati, i differenziali nelle variazioni dei risultati d'interesse nelle aree pilota e di controllo possono essere interpretati come misura dell'impatto della riforma. Tuttavia, se il programma influenza le decisioni d'ingresso o modifica le regole di selezione - e, conseguentemente, le dimensioni e la composizione della popolazione prima e dopo l'introduzione del nuovo regime - il metodo è in grado di misurare la variazione della dimensione del flusso d'ingresso e della sua composizione, ma non può determinare l'impatto sugli altri risultati; questi saranno infatti influenzati anche dalla diversa composizione in sè.

Ad esempio, se i nuovi incentivi e le regole di attivazione di *Pathways* fossero ben noti e influenzassero la decisione di entrare nel programma, le coorti in ingresso nel programma nelle aree pilota e di controllo avrebbero una differente composizione. Per chiarire ulteriormente, si immagini che solo gli uomini entrino nel programma iniziale, e solo le donne entrino una volta che *Pathways* è implementato. Il differente comportamento di uomini e donne risulterebbe mescolato con l'impatto di *Pathways*. In questo caso, sarebbe quindi impossibile separare le differenze nel comportamento attribuibili al genere da quelle determinate dalle differenze nei programmi. Nella pratica, la confusione è più sottile di quella mostrata in questo esempio estremo, ma anche più complessa e insidiosa, perché le popolazioni delle aree pilota e quelle di controllo possono differire anche nelle caratteristiche non osservabili, come la motivazione personale a trovare lavoro, il sostegno ricevuto dalla famiglia e la sensibilità alla modifica delle prestazioni. Per questo motivo, se la valutazione mira a testare l'impatto della riforma su persone simili, i ricercatori devono verificare che la composizione degli ingressi non sia influenzata dall'introduzione del programma.

Nell'esempio di *Pathways*, né la dimensione né la composizione osservabile degli ingressi sono stati influenzati dalla politica. In questo caso, le differenze nella variazione dei risultati tra le aree pilota e di controllo possono essere interpretate come l'impatto delle nuove norme sulla popolazione già presente (e che rimane) nel sistema.

Infine, *Pathways to Work* non introduce variazioni nelle regole d'ingresso ma se ciò fosse successo, le caratteristiche dei beneficiari - prima e dopo la sua introduzione - sarebbero state sistematicamente differenti. Come nell'esempio precedente, l'utilizzo del metodo DiD avrebbe quindi permesso di misurare l'effetto sulla dimensione e la composizione del flusso ma non l'impatto su altre grandezze. Tuttavia, nella misura in cui fosse possibile valutare l'ammissibilità al programma per le persone già inserite nel vecchio programma, si potrebbero confrontare gruppi simili di persone eligibili al programma sia con le vecchie sia con le nuove regole.

Applicabilità alla riforma ipotizzata

In sintesi, l'approccio DiD genera controfattuali a livello di popolazione (la popolazione delle aree pilota e di controllo). Se applicato alla riforma, i valutatori devono

verificare se la politica ha un potenziale impatto sulla dimensione degli ingressi o sulla loro composizione e, se possibile, apportare le modifiche necessarie al disegno di valutazione. Da un lato, se la politica non cambia i flussi d'ingresso, allora è possibile valutare l'impatto delle componenti del sistema sul comportamento di beneficiari simili. Ciò, ovviamente, sempre sotto l'ipotesi di dinamiche parallele tra aree pilota e di controllo in assenza del programma. Dall'altro lato, se la politica incide sugli accessi, la valutazione con il metodo DiD sarà influenzata dalle differenze introdotte nelle dimensioni e nella composizione del flusso. Come avviene nel caso dell'abbinamento statistico, è possibile scomporre i diversi esiti sui beneficiari dei due regimi (per es. sui tassi di occupazione) distinguendo tra quelli indotti dalla variazione della composizione dei gruppi secondo le caratteristiche misurate (età, sesso, istruzione) e un effetto residuo che comprende l'impatto della misura e di tutte le rimanenti caratteristiche non osservabili. Rispetto al caso dell'abbinamento, il metodo della Differenza nelle differenze permette di neutralizzare l'effetto dei diversi contesti economici. Le ipotesi DiD possono però essere indebolite quando sono integrate con l'abbinamento statistico.

In pratica, in questo caso può convenire separare la popolazione in gruppi con livelli di disabilità omogenea ed eseguire l'analisi separatamente per ciascun gruppo (come previsto nella riforma, da "in grado di lavorare" a "nessuna capacità"), confrontando i flussi o gli *stock* di persone nel programma per ciascun gruppo. Ciò richiederebbe di assegnare ciascun disabile a uno dei nuovi tre gruppi definiti dalla riforma.

5.4. Studi controllati randomizzati (RCT)

Descrizione

I due esempi precedenti hanno valutato l'impatto delle riforme attuate in Norvegia e nel Regno Unito senza spiegare l'effetto dei singoli componenti (il caso DiD inoltre ha rilevato che non vi è stato alcun effetto sul flusso). Per determinare quali componenti della riforma sono più efficaci (cioè gli incentivi o l'attivazione e le diverse varianti di ciascuno), e identificare gli individui che beneficiano di più è necessario confrontare ciò che avviene a un insieme di individui esposti a ciascun componente della politica con quanto avviene a un insieme di individui simili non esposti.

Pianificare esperimenti randomizzati piuttosto che valutare le caratteristiche della riforma offre la possibilità di scegliere le domande valutative alle quali rispondere. Ad esempio, un progetto sperimentale nel quale gli individui sono suddivisi a caso in diversi gruppi e a ciascun gruppo vengono offerti diversi elementi del programma può aiutare a spiegare l'impatto dei diversi strumenti di cui questo si compone. Ciò può essere fatto in due modi, sia confrontando persone in luoghi diversi (scegliendo casualmente i distretti ai quali offrire diverse varianti del programma) o persone diverse nello stesso luogo (scegliendo casualmente individui all'interno dello stesso distretto, per esempio in base a giorno e mese di nascita).

Tale confronto può essere fatto mediante aree pilota in ciascuna della quale viene attuata una variante del programma (una generalizzazione del protocollo *Pathways*). Ad esempio, alcune aree possono implementare solo incentivi, altre solo misure di attivazione.

Si deve notare che, quando la partecipazione alle varie componenti del programma è volontaria si potrebbe verificare una distorsione da auto-selezione. In linea di principio, si potrebbe utilizzare l'abbinamento statistico per migliorare la comparabilità. Tuttavia, come precedentemente discusso, in questo caso tale metodo non risulta molto affidabile perché i volontari potrebbero differire sulla base di caratteristiche non osservabili.

Assunzioni

Con questo metodo non è necessario fare affidamento su ipotesi forti come avviene con altri protocolli, perché l'assegnazione casuale a partire da campioni sufficientemente grandi garantisce che gli individui siano simili, in media, sia in termini di caratteristiche osservabili sia di caratteristiche non osservabili. Tuttavia, si deve supporre che le persone non si comportino in modo diverso per il solo fatto di essere consapevoli di partecipare a un esperimento. Questa ipotesi si applica comunque a tutti i tipi di esperimenti, compresi gli studi randomizzati controllati e gli esperimenti naturali, così come i programmi con avviamento graduale[20].

È anche importante che il sistema che si sta valutando sia ben definito e maturo, e quindi simile a quello che sarebbe generalizzato su scala più ampia nella fase a regime. Ancora una volta, questa caratteristica non riguarda solo gli studi randomizzati controllati, in quanto si applica a qualsiasi tipo di valutazione che ha lo scopo di stimare l'impatto di un intervento introdotto gradualmente.

Esempio

> L'autorità danese del mercato del lavoro ha lanciato uno studio randomizzato controllato all'inizio del 2009 per provare su piccola scala alcune disposizioni di un progetto di riforma delle pensioni di invalidità. Rehwald, Rosholm e Rouland (2014) hanno assegnato in modo casuale alcuni lavoratori iscritti ai Centri per l'impiego danesi a un gruppo sperimentale o a un gruppo di controllo. Al gruppo sperimentale sono stati offerti una serie di servizi di attivazione (ritorno graduale al lavoro, azioni di prevenzione sanitaria, etc.) dei quali non potevano invece beneficiare i lavoratori inseriti nel gruppo di controllo. I ricercatori hanno scoperto che i servizi di attivazione non hanno avuto alcun impatto globale, nonostante i loro costi[21].

Applicabilità alla riforma ipotizzata

Un esperimento randomizzato deve essere pianificato in anticipo, prima che la riforma sia attuata. I due elementi principali che possono essere testati sono i servizi di attivazione e gli incentivi finanziari.

20 Gli effetti indotti dalla valutazione (o dall'osservazione) si verificano quando i soggetti modificano il proprio comportamento perché sono consapevoli di partecipare a uno studio e non a causa dell'intervento stesso. Inoltre, anche gli studiosi che effettuano gli esperimenti e misurano gli effetti potrebbero essere prevenuti sapendo di essere parte dello studio: per questo motivo, la raccolta dei dati deve seguire rigorosamente lo stesso protocollo nei gruppi di trattamento e di controllo (cosa che deve comunque avvenire qualsiasi metodo si utilizzi).

21 Questa valutazione è l'unica delle tre discusse in questa sede che misura i risultati sulla salute, non trovando alcun impatto degli interventi testati.

L'attivazione può essere verificata utilizzando l'approccio appena illustrato (Rehwald, Rosholm e Rouland, 2014). Ci sono stati numerosi studi controllati randomizzati sulle politiche di attivazione in diversi paesi europei, tra i quali la Danimarca, la Francia e la Germania.

Va rilevato che l'impatto sul mercato del lavoro (il più importante), va osservato per un periodo prolungato, il che implica che i due gruppi (sperimentale e di controllo) devono rimanere separati per periodi prolungati. Questi risultati possono essere infatti misurati solo una volta che la maggiore motivazione al lavoro si è tradotta in un lavoro vero e proprio, un processo che potrebbe richiedere un certo tempo. Ciò richiede che l'esperimento continui e che la generalizzazione dei risultati sia sospesa fino a che essi non siano realmente osservabili[22].

Gli incentivi finanziari al lavoro possono essere valutati in modo simile. Ci sono stati diversi esperimenti randomizzati in Canada e negli Stati Uniti su diverse popolazioni (si veda il prossimo capitolo sulla riforma del reddito minimo garantito).

Le disposizioni da inserire in un processo di riforma potrebbero essere testate preventivamente. I sussidi all'occupazione potrebbero essere modificati, ad esempio aumentando l'importo della sovvenzione accorciandone la durata, o concentrandosi sul lavoro che la persona disabile svolgeva prima della sua disabilità. Le sovvenzioni potrebbero anche promuovere l'adattamento del posto di lavoro alla disabilità del lavoratore. In questo caso si potrebbero abbinare casualmente le persone a differenti servizi, oppure abbinare i diversi servizi a diversi distretti. Tali esperimenti potrebbero anche essere effettuati mentre lo schema di riforma di base è già in vigore.

5.5. Progetti pilota e RCT

Se per motivi legali o amministrativi non è fattibile assegnare persone diverse a protocolli in vigore e riformati allo stesso tempo e nello stesso posto, la riforma potrebbe essere introdotta gradualmente per aree geografiche, in modo che tutti i residenti nella stessa zona abbiano le stesse regole e incentivi. Nell'ambito di questo regime, i gruppi sperimentali (riforma) e di controllo (disposizioni vigenti) sarebbe composti da aree geografiche. Ciò equivarrebbe a utilizzare esattamente lo stesso tipo di valutazione Differenza nelle differenze del programma *Pathways to Work*, ma le aree pilota dove la riforma inizia a essere implementata sarebbero scelte a caso (le altre aree servirebbero come controllo), piuttosto che da parte dell'amministrazione. L'assegnazione casuale aumenta la comparabilità delle aree sperimentali e di controllo senza ricorrere all'assunzione delle "dinamiche parallele". Tale randomizzazione a livello di area è stata realizzata, per esempio, in Francia, per valutare l'assistenza nella ricerca del lavoro (Crépon *et al.*, 2013) una volta decisa la sua introduzione graduale.

L'assegnazione casuale a volte si scontra con dei vincoli politici. Ad esempio, i governi potrebbero voler selezionare aree altamente performanti come piloti per promuovere una riforma, o al contrario, concentrarsi su settori poco performanti in modo da metterli alla prova. Nel caso di *Pathways*, il governo ha scelto le aree pilota dove i

22 Ciò è richiesto da qualsiasi protocollo di valutazione che misura i risultati sul mercato del lavoro: è necessario attendere i risultati della valutazione prima di prendere una decisione politica.

Jobcentre Plus hanno lavorato bene e per un lasso di tempo sufficiente. In pratica, la definizione delle aree potrebbe dover seguire confini amministrativi.

Detto questo, l'assegnazione casuale di aree pilota fornisce risultati più affidabili rispetto a una selezione intenzionale quando una riforma è introdotta gradualmente. Se l'assegnazione di aree pilota e di controllo non è casuale, "per costruzione" le aree pilota non possono essere direttamente comparabili con quelle di controllo. In questo caso, possono essere invocate assunzioni di "dinamiche parallele", ma questa rimane un'ipotesi non verificabile. Inoltre, quando l'assegnazione non è casuale e zone tipiche sono scelte come pilota (il meglio, o il peggio), la validità esterna della valutazione risulta limitata. Ci può essere un equilibrio tra il definire un vasto insieme di aree abbastanza testando il programma su un sottoinsieme randomizzato di esse, oppure scegliere le migliori; in ogni caso, tutti i risultati sarebbero applicabili alle sole aree "mature".

6. Requisiti istituzionali, organizzativi e politici

Requisiti istituzionali, organizzativi e politici possono anche influenzare le decisioni metodologiche, come dimostra il caso della valutazione di *Pathways*.

Il programma *Pathways to Work* doveva essere attuato nei Jobcentre Plus, un nuovo tipo di servizio pubblico per l'impiego risultante dalla fusione del Servizio per l'occupazione e del Dipartimento per la Sicurezza Sociale. Quando i primi *Pathways* pilota sono stati avviati, nel mese di ottobre del 2003, solo un terzo degli uffici di collocamento utilizzava già il modello Jobcentre Plus; ciò restrinse significativamente la possibilità di selezionare le aree pilota, influenzando quindi il disegno della valutazione. Il fatto che, nel mese di ottobre 2003, il Dipartimento per il Lavoro e le pensioni (DWP) avesse già messo in esecuzione sei progetti pilota *welfare-to-work* in tutto il Regno Unito (mettendo il proprio *staff* sotto sforzo), complicò ulteriormente la selezione dei siti pilota.

Problemi di costo e di capacità

La valutazione di *Pathways* ha considerato l'impatto del programma nel suo insieme. In altre parole, non ha messo in luce se un particolare componente del pacchetto (ad esempio, le interviste obbligatorie, il credito per il ritorno al lavoro, etc.) abbia avuto più importanza di altri nel determinarne l'impatto complessivo. I ricercatori hanno lamentato che la valutazione non fosse stata progettata per dare un quadro più completo dell'efficacia della politica (Adam *et al.*, 2006: 4).

Tuttavia, i funzionari del Dipartimento per il lavoro e le pensioni (DWP) hanno sottolineato che la progettazione di una valutazione in grado di misurare l'impatto di ciascuna delle diverse componenti di *Pathways* avrebbe richiesto una sperimentazione più estesa, complessa e costosa, esponendoli anche al rischio di fornire risultati inconcludenti (Boa *et al.*, 2010: 22).

Riferimenti bibliografici

Adam S., Emmerson C., Frayne C., Goodman A. (2006), Early quantitative evidence on the impact of the Pathways to Work pilots. Institute for Fiscal Studies and

Department for Work and Pensions, Research Report 354. Norwich: Stationery Office.

Autor D.H., Duggan M.G. (2003), The Rise in the Disability Rolls and the Decline in Unemployment. *The Quarterly Journal of Economics*, 118, 1: 157-205.

Boa I., Johnson P., King S. (2010), The impact of research on the policy process. Frontier Economics Ltd and Department for Work and Pensions, Working Paper n. 82. Norwich: Stationery Office.

Crépon B., Duflo E., Gurgand M., Rathelot R., Zamora P. (2013), Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. *The Quarterly Journal of Economics*, 128, 2: 531-580.

Kai R., Rosholm M., Rouland B. (2013), Does Activating Sick-Listed Workers Work? Evidence from a Randomized Experiment. (work in progress).

OECD (2009), Pathways onto (and off) Disability Benefits: Assessing the Role of Policy and Individual Circumstances. In: *OECD Employment Outlook 2009 -Tackling the Jobs Crisis*. Paris: OECD. Chapter 4. ISBN 978-92-64-06791-2.

Ravndal K.A., Mogstad M. (2014), How Financial Incentives Induce Disability Insurance Recipients to Return to Work. American Economic Review, 104, 2: 624–655.

Rehwald K., Rosholm M., Rouland B., (2014), Does Activating Sick-Listed Workers Work? Evidence from a Randomized Experiment. Mimeo.

Stuart A., Bozio A., Emmerson C. (2009), Can we estimate the impact of the Choices package in Pathways to Work? Norwich, UK: The Stationery Office, Department for Work and Pensions, Working paper n. 60.

Stuart A., Bozio A., Emmerson C. (2010), Reforming Disability Insurance in the UK: Evaluation of the Pathways to Work Programme. London: Institute for Fiscal Studies.

ESEMPIO 2

ESEMPIO 2 - COME VALUTARE UNA RIFORMA DEL REDDITO MINIMO GARANTITO

1. Introduzione

Questo caso di studio mostra come valutare l'impatto dei sistemi di reddito minimo. Sono presentati i principali disegni di ricerca disponibili e si mostra come siano stati utilizzati in passato. Tali disegni si distinguono essenzialmente per: i) i requisiti metodologici; ii) le assunzioni relative alla comparabilità dei gruppi sperimentali e di controllo.

La discussione è illustrata con esempi tratti dagli Stati Uniti, Canada, Cipro, Francia e Regno Unito. Il caso di studio è organizzato come segue: la sezione 2 fornisce una panoramica dei sistemi di reddito minimo; la sezione 3 descrive brevemente le caratteristiche di una probabile riforma da prendere a riferimento; la sezione 4 spiega come costruire situazioni controfattuali e discute i diversi metodi che potrebbero essere applicati per accompagnare la riforma con un idoneo disegno di valutazione. La discussione è illustrata da esempi concreti.

2. Cosa sono i programmi di Reddito minimo garantito

Quasi tutti i paesi europei hanno stabilito programmi di reddito minimo volti principalmente a ridurre la povertà. Questi trasferimenti, in contanti o in natura, mirano a fornire un adeguato *standard* di vita alle famiglie che non dispongono di un reddito sufficiente. Spesso funzionano come rete di sicurezza di ultima istanza, insieme ai sussidi di disoccupazione ma, in alcuni paesi, costituiscono il principale strumento di protezione sociale. I sistemi di reddito minimo includono, in genere, prestazioni di assistenza indipendenti dallo status occupazionale, dai contributi versati, dai benefici per genitori soli, sussidi per l'alloggio e crediti d'imposta.

Una sfida politica importante consiste nel progettare un sistema che garantisca un reddito minimo per coloro che non possono permettersi un tenore di vita accettabile senza costituire un disincentivo al lavoro. Se l'importo delle prestazioni erogate è infatti superiore ai guadagni attesi dal lavoro, il reddito minimo può disincentivare

al lavoro, in toto (facendo affidamento completo al sussidio) o in parte (riducendo il numero di ore lavorate). Se, al contrario, l'importo delle prestazioni è troppo basso, o le misure di attivazione non sono adeguatamente focalizzate, le famiglie eligibili ma non in grado di lavorare non sarebbero adeguatamente protette contro la povertà e la miseria. Su questo equilibrio possono inoltre influire limiti temporali e altre condizioni specifiche.

Le politiche attive del lavoro e le integrazioni del reddito sono due alternative politiche promosse per incoraggiare i beneficiari dei servizi di welfare a lavorare mantenendo un'adeguata rete di sicurezza. Capire se queste politiche sono in grado di creare gli incentivi appropriati per tornare al lavoro, garantendo al contempo un adeguato tenore di vita è una questione empirica.

3. Una riforma del reddito minimo garantito

Un regime di reddito minimo può variare in base a:

- la generosità dei benefici e il loro profilo;
- le misure disponibili (servizi di sostegno all'infanzia, assistenza sanitaria, sostegno all'alloggio, indennità di istruzione);
- i criteri di ammissibilità;
- la condizionalità comportamentale (la necessità di adempiere alle misure di attivazione);
- il destinatario della misura (l'individuo o la famiglia).

Una riforma del reddito mimino potrebbe quindi mirare a:

- evitare le duplicazioni, sostituendo altre misure di assistenza con un sistema di gestione accentrata del reddito minimo;
- migliorare gli incentivi al lavoro, implementando requisiti stringenti per conformarsi alle politiche attive del lavoro vigenti;
- rafforzare il monitoraggio dei requisiti supplementari (famiglia, disabilità, benefici per la salute e borse di studio).

Il reddito minimo garantito può comprendere una quota di base, un'indennità di alloggio, uno sgravio fiscale, così come un assegno *una tantum* previsto nel caso di esigenze straordinarie. Il beneficiario è di solito la famiglia e i principali criteri di ammissibilità sono che i bisogni di base della famiglia superino il suo reddito e la necessità di adempiere alle condizioni di attivazione. Il regime coprirebbe un gruppo eterogeneo di beneficiari, comprese le famiglie che hanno esaurito l'indennità di disoccupazione, quelle che non risultano eligibili ai sussidi di disoccupazione e le famiglie dei lavoratori i cui redditi non possono coprire i bisogni fondamentali.

In generale, i principali **risultati** su cui valutare una riforma di questo tipo sono:

- i livelli di povertà: variazione del reddito e dei consumi netti;
- l'offerta di lavoro, in termini di partecipazione al mercato del lavoro e di ore lavorate (del principale beneficiario e/o del coniuge);
- la progressione salariale;

- distribuzione intra-famigliare del reddito (ad esempio l'impatto sul benessere del coniuge, impatto sul benessere dei bambini).

I **meccanismi** della riforma dipendono da vari aspetti, i più importanti sono:

- le modifiche nel processo di verifica dei mezzi dei potenziali beneficiari;
- la sensibilità dei potenziali beneficiari del reddito minimo garantito ai trasferimenti (rispetto al salario);
- l'influenza delle misure supplementari sui redditi familiari e sugli incentivi all'occupazione;
- l'influenza delle politiche di attivazione sull'adesione allo schema di reddito minimo (efficacia e durata).

4. Come costruire controfattuali

Per valutare se la riforma fa la differenza, quanto e con quali costi e benefici, si ha la necessità di stabilire una situazione controfattuale[23], che è una stima di ciò che sarebbe successo alle grandezze di interesse in assenza della riforma.

La riforma del reddito minimo modifica i criteri di ammissibilità al sistema, così come il livello dei diritti e delle misure di attivazione previste. Di conseguenza, la riforma può potenzialmente cambiare la dimensione e la composizione del flusso di accesso allo schema (risultati a livello di popolazione), nonché i tassi di partecipazione al mercato del lavoro e dei livelli di reddito dei beneficiari (risultati a livello individuale).

Al fine di misurare l'impatto della riforma sui risultati a livello individuale (cioè la partecipazione al mercato del lavoro, il reddito netto, la durata delle prestazioni) si ha la necessità di costruire controfattuali a livello individuale. In questo caso, si tratta di mettere a confronto i beneficiari del nuovo regime con beneficiari simili del programma ante riforma. Controfattuali a livello individuale possono essere utilizzati anche per stimare l'impatto dei diversi aspetti della riforma e individuare quelli più efficaci. Questo è di grande valore dal punto di vista politico, in quanto la spesa pubblica può essere ottimizzata, destinandola alle opzioni politiche che dimostrano un impatto maggiore sui risultati desiderati. Inoltre, controfattuali a livello individuale aiutano a stimare l'eterogeneità dell'impatto su diverse sotto-popolazioni consentendo di costruire azioni su misura. Ciò è molto importante per politiche che mirano ad attivare i destinatari degli assegni di reddito minimo i quali, oltre a costituire un gruppo molto eterogeneo, in genere hanno molte difficoltà a trovare un'occupazione (ad esempio rispetto ai beneficiari degli assegni di disoccupazione).

Al fine di valutare l'impatto della riforma sugli ingressi nel sistema, è necessario costruire controfattuali a livello di popolazione. In questo caso sarebbero comparate due popolazioni simili: quella che accede al regime iniziale e quella che accede al nuovo regime di reddito minimo. La differenza nei tassi di adesione, nell'iscrizione e nelle caratteristiche delle famiglie riscontrabili tra le popolazioni che accedono al programma iniziale di assistenza e al nuovo regime potrebbe dare una misura delle

23 Si veda a proposito; ESF Guide on Counterfactual Evaluation: Design and Commissioning of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2012.

variazioni nelle dimensioni e nella composizione del flusso in ingresso. Tuttavia, questo tipo di analisi potrebbe essere difficilmente praticabile se il sistema iniziale risulta frammentato in diversi programmi di assistenza sociale, amministrati da diversi ministeri e dipartimenti. Le informazioni raccolte su base decentrata potrebbero anche fornire dati contrastanti sulla partecipazione ai vari programmi di assistenza sociale.

Si deve notare che, quando la riforma modifica gli accessi al nuovo regime, i risultati a livello individuale possono combinare effetti derivanti sia dalla composizione dei flussi che dai contenuti del nuovo servizio. Infatti, se la riforma cambia i criteri di ammissibilità, le persone che decidono di aderire al nuovo regime di reddito minimo sono probabilmente diverse da quelli che avrebbero aderito al regime iniziale, ostacolandone la comparabilità. Più precisamente, quando la composizione dei flussi viene modificata, qualsiasi modifica nei risultati a livello individuale (la partecipazione al mercato del lavoro, il reddito netto e la durata delle prestazioni) potrebbe essere indotta simultaneamente sia dalle specifiche caratteristiche dei beneficiari del reddito minimo garantito che hanno avuto accesso alle prestazioni in base alle nuove norme di ammissibilità, sia dalle condizioni del nuovo servizio. Al fine di distinguere la parte di impatto dovuta alla variazione delle caratteristiche individuali da quella attribuibile ai meccanismi della riforma, si possono identificare e confrontare gli individui più simili che accedono ai due regimi. Questo può essere fatto solo se si accetta l'ipotesi che le caratteristiche osservate sono sufficienti a rendere gli individui comparabili (si veda oltre). Come nel caso della riforma dell'assicurazione di invalidità, risulta molto difficile misurare contemporaneamente variazioni dei flussi (risultati a livello di popolazione) e cambiamenti nello status occupazionale e nei livelli di reddito (risultati a livello individuale).

5. Potenziale dei diversi metodi di valutazione d'impatto controfattuale

5.1. Abbinamento statistico

Descrizione

Con questo metodo si costruisce un controfattuale abbinando partecipanti e non partecipanti al programma in base a una serie di caratteristiche osservabili. Un abbinamento di successo richiede una ricerca preliminare allo scopo di identificare le variabili che potrebbero essere statisticamente correlate alla probabilità di partecipare al programma e ai suoi risultati. Per creare corrispondenze sufficienti, sono necessari campioni estesi. Questo metodo è in grado di fornire una stima dell'effetto di un intervento per tutti i partecipanti che possono essere abbinati con successo a un non partecipante.

Assunzioni

L'ipotesi principale è che tutte le caratteristiche di base che influenzano partecipazione e risultati d'interesse possano essere osservate e valutate.

Applicabilità alla riforma ipotizzata

In teoria, questo metodo potrebbe essere utilizzato per misurare sia risultati a livello individuale sia i cambiamenti negli ingressi al nuovo regime.

Allo scopo di valutare l'impatto della riforma a livello individuale, si potrebbero abbinare i beneficiari del precedente regime a quelli del nuovo schema, in base a un insieme di caratteristiche osservate e poi confrontare gli esiti sui due gruppi (lo status occupazionale, il livello di reddito, etc.). Tra le caratteristiche dovrebbero essere inclusi i criteri di ammissibilità del vecchio e del nuovo programma; in modo da identificare le famiglie non ammissibili al nuovo regime che andrebbero valutate separatamente. Un uso alternativo dell'abbinamento potrebbe essere quello di confrontare l'effetto di varie forme di incentivi al lavoro (o la loro assenza) sui partecipanti al nuovo schema, senza confrontare il vecchio e il nuovo schema.

Per valutare l'effetto sugli ingressi si dovrebbero abbinare i potenziali beneficiari, alcuni dei quali si sono confrontati con il regime iniziale e altri con quello nuovo, e confrontare la loro probabilità di ingresso nel programma di reddito minimo.

Si noti tuttavia, che in questo contesto, il metodo dell'abbinamento statistico presenta lacune importanti:

- Come già accennato in precedenza, questo metodo richiede che tutte le caratteristiche determinanti per la partecipazione e i risultati di interesse siano osservabili e rappresentabili. Sui tratta di un'ipotesi forte e, in questo caso particolare, altamente improbabile. È importante notare che la decisione di partecipare al programma è volontaria, pertanto è difficile determinare in che misura essa sia determinata dalle caratteristiche del nuovo schema di reddito minimo, e quanto da differenze pre-esistenti nelle caratteristiche osservabili e non osservabili degli individui che hanno scelto di partecipare.
- Inoltre, se il nuovo reddito minimo sostituisce completamente i programmi di assistenza precedenti e il confronto si basa su dati retrospettivi, i partecipanti al nuovo

programma non saranno più comparabili ai partecipanti delle vecchie misure, in quanto saranno probabilmente immersi in una differente situazione economica.

5.2. Confronto attorno al punto di discontinuità (RDD)

Descrizione

Questo metodo confronta individui posizionati appena al di sopra di una determinata soglia di ammissibilità continua, con quelli posizionati appena sotto. Si tratta probabilmente di individui molto simili e la soglia determina se siano o meno esposti all'intervento che si intende valutare. L'ampiezza dell'intervallo attorno alla soglia determina la dimensione del campione.

Assunzioni

Questo metodo si basa sul presupposto che l'intervento si basi su un criterio di selezione chiaramente quantificabile basato su un punteggio continuo e che i partecipanti non possono influenzare i punteggi attorno alla soglia. Inoltre, si assume che gli individui appena sotto e appena sopra la soglia non siano significativamente diversi.

Esempio

Jones (2013) ha valutato l'impatto della Earned Income Tax Credit (EITC) sul numero di ore lavorate attraverso un Regression Kink Design, una variante del confronto attorno al punto di discontinuità (RDD)[24]. L'EITC è un programma introdotto nel 1975 negli Stati Uniti che offre la possibilità di ottenere un credito d'imposta sulla tassazione del reddito, alle persone a basso reddito e alle coppie. Esso mira a trasferire reddito alle famiglie a basso reddito e, allo stesso tempo, a incoraggiare e sostenere coloro che scelgono di lavorare. L'ammissibilità dipende da tre criteri principali: il contribuente deve avere un reddito positivo, tale reddito non deve superare una determinata soglia, e anche se i contribuenti senza figli hanno diritto a un piccolo credito, i vantaggi più significativi sono riservati ai contribuenti con figli conviventi.

I sostenitori del programma sostengono che il credito d'imposta per i più bisognosi incentivi la partecipazione al mercato del lavoro, perché il credito è accessibile solo ai contribuenti. Tuttavia, non è chiaro come la struttura del credito possa incentivare, oltre alla partecipazione, anche il numero di ore lavorate. In una fase iniziale, infatti, il credito cresce al crescere del reddito, nella fase successiva il credito rimane costante al crescere del reddito, per poi diminuire gradualmente fino ad annullarsi. I contribuenti nella fase di inserimento si confrontano quindi con un effetto di sostituzione positivo, in quanto il credito aumenta con il numero di ore lavorate. I contribuenti che si trovano invece nella parte di curva piatta trovano un effetto reddito negativo, perché il programma offre una quantità costante di credito a prescindere dal numero di ore lavorate. In questo scenario, il contribuente può quindi raggiungere un determinato livello di utilità lavorando meno ore di quanto sarebbe richiesto in assenza del programma. Ciò significa che, all'aumentare del reddito, i contribuenti "comprano" più tempo libero, riducendo il numero di ore lavorate (il tempo libero è un bene come un altro). Infine, i contribuenti posizionati nella fase decrescente hanno a che fare con un effetto di sostituzione negativo e un effetto reddito negativo. L'ammontare del credito dimi-

24 Tale variante, denominata RKD, è stata proposta da Card *et al.* (2012).

nuisce all'aumentare del numero di ore lavorate, rendendo un'ora extra di svago relativamente meno costosa di un'ora in più di lavoro (effetto di sostituzione negativo). Inoltre, all'aumentare del reddito, i contribuenti possono "comprare" più tempo libero, riducendo il numero di ore lavorate (effetto reddito negativo).

Analogamente al metodo RDD, il metodo RKD si basa su un punto di svolta (nodo) della regola alla base della politica per identificarne l'effetto causale. In questo caso, l'autore si avvale delle discontinuità nella funzione di beneficio del programma EITC per esaminare come queste influenzino il numero di ore lavorate dalle madri. L'importo delle prestazioni ricevute è, infatti, una funzione del reddito. Questa funzione è continua, ad eccezione di due punti o "nodi" (poco prima di entrare nella fase "piatta" della funzione, e al termine della fase piatta, quando inizia la graduale eliminazione del beneficio). L'autore ha quindi confrontato il numero di ore lavorate dai beneficiari che si trovano appena prima di un nodo, con quelle di coloro che si trovano appena dopo tale nodo. I risultati hanno mostrato che le madri *single* adeguano il proprio comportamento in modo da massimizzare i benefici. In questo modo, le madri con più di un bambino riducono il numero di ore lavorate quando il loro reddito nel periodo precedente è diminuito perché si sono venute a trovare subito dopo il nodo nella funzione di beneficio del programma, dove il vantaggio inizia a diminuire.

La prima ipotesi è che i due gruppi di donne nei pressi del "nodo" siano simili nelle caratteristiche osservate e non osservate. Potrebbero esservi notevoli differenze nel caso in cui, per esempio, lo stesso punto di soglia sia stato usato per fornire altri tipi di servizi in grado di influenzare i risultati da osservare. Ciò si verificherebbe nel caso che anche altre misure fiscali e di trasferimento cambiassero in prossimità delle stesse soglie dell'EITC. L'autore sottolinea che le donne con un solo figlio affrontano diversi oneri e crediti (assegni familiari) federali a seconda del lato del nodo nel quale si trovano. Allo stesso modo (a seconda del livello del reddito e quindi del lato del "nodo" in cui sono posizionate), le donne con un figlio o più sono soggette anche ad aliquote marginali differenti delle imposte statali. Per questi gruppi, il metodo RKD non consentirebbe di distinguere l'effetto della variazione di imposte e crediti diversi dagli incentivi dell'EITC.

La seconda assunzione è che, anche se le donne possono modificare il numero di ore lavorate, una volta che hanno scoperto la posizione nella quale si sono venute a trovare l'anno precedente, non sono in grado di prevedere la posizione reddituale e fiscale nella quale si troveranno nell'anno in corso. L'autore sostiene che tale capacità di previsione è improbabile, in quanto i punti di soglia cambiano ogni anno. L'autore fornisce alcune prove a supporto, mostrando che i redditi non sono concentrati attorno ai punti nodali.

Applicabilità alla Riforma

Il metodo RDD può essere applicato alla riforma se è disponibile qualche variabile continua che determina l'ingresso nel nuovo regime di reddito minimo. Un possibile candidato potrebbe essere la data di attuazione della misura. I candidati che entrano nel programma di reddito minimo dopo una certa data verranno assegnati al nuovo schema, mentre quelli entrati prima rimarrebbero nelle condizioni iniziali del programma di assistenza sociale. La data di soglia deve essere determinata retroattivamente, in modo che i beneficiari del reddito minimo non possano modficare le proprie decisioni di entrata nel nuovo regime. Inoltre, il programma di assistenza sociale iniziale deve procedere in parallelo con il nuovo regime di reddito minimo (almeno durante la fase della sua valutazione), in modo che gli individui inseriti in ciascun programma possano essere osservati durante lo stesso periodo di tempo. Un altro approccio potrebbe basarsi su una soglia di reddito che determini l'ammissibilità al programma, in quanto gli individui che si posizionano in prossimità dei due

lati di tale soglia possono essere considerati molto simili. Tuttavia, questo approccio non sarebbe valido se gli individui potessero modificare il proprio reddito in modo da risultare ammissibili. Purtroppo, ciò risulta effettivamente probabile nella maggior parte dei contesti istituzionali.

Con questo metodo è possibile misurare l'impatto complessivo del nuovo programma, senza però distinguere l'effetto provocato dai nuovi livelli dei benefici da quello delle misure di attivazione.

Inoltre, il metodo RDD non può aiutare nella stima degli impatti sugli ingressi e richiede flussi consistenti; per poter garantire la potenza statistica necessaria per rilevare un impatto è infatti necessario che almeno alcune centinaia di individui si trovino vicini al punto di discontinuità.

5.3. Differenza nelle differenze (DID)

Descrizione

Il metodo confronta nel tempo la variazione dei risultati ottenuti prima e dopo l'inizio del programma tra un gruppo di partecipanti e non partecipanti. Fornisce una misura dell'impatto sull'intera popolazione dei partecipanti, controllando per condizioni costanti (osservate o meno) che possono essere correlate sia con i risultati finali sia con il fatto di essere parte del gruppo di controllo.

Con la Differenza nelle differenze è possibile sia confrontare un gruppo di persone che ha i requisiti per ricevere l'intervento con un gruppo simile ma non ammissibile, sia confrontare aree pilota nelle quali viene introdotto il programma con aree di confronto che non lo ricevono.

Assunzioni

L'approccio è basato sull'assunzione delle "dinamiche parallele". Al fine di determinare se la differenza nei risultati è dovuta al programma, si deve infatti assumere che le dinamiche dei risultati di partecipanti e non partecipanti rimarrebbero uguali in assenza del programma e che, in tal caso, rimarrebbe invariata anche la composizione di ciascun gruppo. Un modo per validare l'ipotesi è verificare se entrambi i gruppi hanno mostrato andamenti paralleli prima dell'introduzione del programma. Altre modalità di verifica prevedono, da un lato l'esecuzione di test che simulano trattamenti "placebo" (assegnati casualmente a un sottoinsieme delle unità dei gruppi di intervento e controllo) oppure stimano gli effetti su risultati che non dovrebbero essere ragionevolmente attribuibili all'intervento, dall'altro utilizzano gruppi di controllo differenti.

Esempio

L'impatto delle riforme EITC negli Stati Uniti sono stati ampiamente studiati tramite DID. Eissa e Liebman (1996) hanno utilizzato questo metodo per valutare

l'impatto dell'espansione del programma EITC del 1987[25]. L'obiettivo della politica era l'aumento della partecipazione al mercato del lavoro e gli orari di lavoro delle donne con figli. Gli autori si sono concentrati sulle donne sole con figli, che sono il gruppo più numeroso dei contribuenti che possono beneficiare della EITC. Al momento della valutazione, uno dei criteri di ammissibilità era avere almeno un figlio convivente. Utilizzando una strategia di valutazione basata sulla Differenza nelle differenze, gli autori hanno quindi confrontato la variazione dell'offerta di lavoro delle donne sole con figli (gruppo di intervento, potenzialmente ammissibile per l'EITC) prima e dopo l'estensione dell'EITC con quella delle donne single senza figli (gruppo di controllo, non ammissibili per il EITC). In questo modo si è scoperto che l'offerta di lavoro delle donne sole con figli è aumentata più di quello delle donne senza figli. Non è stato invece trovato alcun impatto sulla quantità di ore lavorate.

Il metodo delle Differenze nelle differenze aiuta a controllare i fattori ambientali (nuove politiche, congiuntura economica) che potrebbero indurre un cambiamento dell'offerta di lavoro. I valutatori utilizzano inoltre diversi gruppi di controllo per supportare la loro strategia di valutazione. Tuttavia, i valutatori si basano su due importanti presupposti. In primo luogo, essi ipotizzano che le donne sole con figli si sarebbero comportate in modo simile alle donne single senza figli in assenza dell'estensione di EITC. Essi forniscono alcuni elementi di supporto all'ipotesi mostrando che le tendenze di lungo periodo di partecipazione al mercato del lavoro delle donne non sono molto diverse, anche se la partecipazione delle madri sole sembra essere più sensibile al ciclo economico.

Il secondo fondamentale presupposto è che, oltre all'estensione di EITC, non si siano riscontrati altri *shock* che potrebbero aver influenzato in modo differenziale i risultati dei gruppi d'intervento e di controllo. Ciò sarebbe potuto avvenire se si fosse verificato un cambiamento in altre politiche fiscali e creditizie, nella dinamica del ciclo economico o a causa altri *shock* economici che colpiscono in modo differente le madri e le donne sole senza figli. Questa è un'ipotesi molto forte, in quanto è molto difficile escludere l'esistenza di *shock* sconosciuti.

Blundell, Brewer e Shephard (2005) hanno utilizzato una strategia simile per misurare l'impatto del Working Families' Tax Credit (WFTC), introdotto in Gran Bretagna nel mese di ottobre 1999. L'obiettivo di questo progetto è stato quello di sostenere le famiglie dei lavoratori a basso reddito con figli. Gli autori hanno confrontato i tassi di occupazione dei genitori con quelli dei non-genitori, partendo dal presupposto che le tendenze occupazionali sottostanti avrebbero seguito percorsi simili in assenza del programma. L'esito dello studio mostra che il WFTC e le altre contestuali riforme dei regimi fiscali e delle prestazioni hanno aumentato i tassi di occupazione delle madri sole e diminuito quelli dei padri con partner.

Un altro esempio è la valutazione del programma francese "Revenu de Solidarité Active" (RSA). Il RSA è stato introdotto nel 2009 in sostituzione di altri schemi di assistenza sociale. Il nuovo regime prevede maggiori incentivi per il rientro al lavoro attraverso trasferimenti monetari condizionati all'occupazione. L'importo delle prestazioni è aumentato dopo un anno di lavoro, al fine di incoraggiare la stabilità del lavoro. Oltre a ciò, la durata delle prestazioni è stata estesa e il servizio di consulenza al lavoro è stato rinforzato.

25 L'espansione dell'EITC (Earned Income Tax Credit – Credito d'Imposta sul Reddito Imponibile) fu il risultato dalla legge di riforma fiscale del 1986 e si sostanziò in un aumento del tasso di sovvenzione per i livelli di reddito vicini al limite inferiore di applicazione dell'agevolazione fiscale, un aumento del credito massimo e una riduzione del livello di reddito massimo che consentiva di ottenere l'agevolazione.

La valutazione ha introdotto una dimensione sperimentale, stabilendo aree pilota e di controllo. I principali risultati che sono stati presi in considerazione sono i tassi di occupazione e la qualità del lavoro. Le aree pilota sono state scelte dal governo e i valutatori hanno suggerito quelle di controllo, abbinandole sulla base di una serie di criteri socio-demografici.

In questa sede è importante notare che, anche se da un punto di vista gestionale è concepibile selezionare aree pilota "atipiche" - o perché mostrano un rendimento migliore della media (ad esempio, per sostenere una riforma) o peggiore rispetto alla media (per contestare una riforma) - questa scelta non casuale può minare la comparabilità e, potenzialmente, produrre risultati distorti. A parità di altre condizioni, la selezione di un'area particolarmente virtuosa come sito pilota potrebbe aumentare la probabilità di osservare un risultato positivo. Al contrario, le aree a basso rendimento potrebbero aumentare le probabilità di osservare esiti negativi. In ogni caso, selezionare le aree pilota perché hanno prestazioni particolarmente positive o negative riduce la validità esterna della valutazione. Se la politica verrà implementata in aree con *performance* diverse, l'impatto del programma a regime sarà probabilmente diverso da quello del pilota.

Sia la valutazione dell'EITC sia quella della RSA devono affrontare la sfida di una scarsa potenza statistica. È molto difficile identificare con certezza i potenziali beneficiari del reddito minimo, cosicché entrambe le valutazioni hanno dovuto misurare i risultati rilevanti (occupazione, reddito) su campioni molto grandi senza sapere esattamente quali individui fossero in una situazione occupazionale che rendeva più probabile il loro ingresso nel programma a regime o li rendeva sensibili a variazioni delle sue caratteristiche. Il campione dell'EITC comprendeva tutte le donne single, mentre

nel campione RSA erano inclusi tutti i beneficiari di due sistemi di welfare il "RMI"[26] o l'API[27].

Applicabilità alla riforma ipotizzata

Il metodo della Differenza nelle differenze rende possibile misurare l'impatto di un programma di reddito minimo sui flussi di ingresso o su altre variabili come la partecipazione al lavoro e i livelli di reddito. Da una parte, se la politica non cambia gli accessi alla misura, questo disegno può valutare l'impatto degli elementi del regime sul comportamento di beneficiari simili. Questo è vero sempre sotto l'ipotesi che le tendenze nei risultati sui gruppi di intervento e di controllo sarebbero gli stessi in assenza del programma.

D'altra parte, se la politica incide sugli ingressi, la valutazione con questo metodo sarà influenzato dalle differenze nei flussi e nella loro composizione. Come avviene nel caso dell'abbinamento statistico, è possibile scomporre i diversi esiti sui beneficiari dei due regimi (per es. sui tassi di occupazione) tra quelli indotti dalla variazione della composizione dei gruppi secondo le caratteristiche misurate (età, sesso, istruzione) e un effetto residuo che comprende l'impatto della misura e di tutte le rimanenti caratteristiche non osservabili. Rispetto al caso dell'abbinamento, la configurazione del metodo della Differenza nelle differenze permette di neutralizzare l'effetto di diversi contesti economici.

Per valutare questa particolare riforma, il metodo può essere implementato soprattutto in due casi. In primo luogo se alcuni gruppi predeterminati sono esclusi dal programma o affrontano sue versioni differenti. Questi gruppi devono essere scelti sulla base di variabili oggettive, come l'età, la dimensione della famiglia, e la situazione occupazionale. In tal caso, si possono seguire i gruppi nel tempo e assumere che, in assenza della riforma, il loro impiego o la loro situazione economica si sarebbero evolute in modo parallelo. Un'altra applicazione del metodo può derivare da una implementazione progressiva della politica, per cui alcune zone siano scelte come pilota e altre come controllo, come nell'esperimento Francese della RSA.

5.4. Studi controllati randomizzati (RCT)

Descrizione

Gli studi controllati randomizzati misurano l'impatto medio di una politica o di un programma assegnando casualmente i soggetti ai gruppi di intervento e di controllo e confrontando la differenza nei risultati ottenuti.

Assunzioni

In questo caso non è necessario fare affidamento su ipotesi forti come richiesto da altri protocolli, perché l'assegnazione casuale da campioni sufficientemente grandi garantisce che gli individui siano simili, in media, sia per quanto riguarda le caratteristiche osservabili sia per quelle non osservabili. L'unica assunzione necessaria è che

26 Revenu Minimum d´Insertion (Reddito minimo di inserimento).

27 Allocation de Parent Isolé (Assegno per genitore solo).

le persone non si comporteranno in modo diverso perché consapevoli di essere inserite in un esperimento. Questa ipotesi si applica comunque a tutti i tipi di esperimento.

È anche importante che il sistema in corso di valutazione sia ben definito e maturo e, quindi, simile a quello potrebbe essere effettivamente generalizzato a scala più ampia. Anche questo aspetto non riguarda solo gli studi randomizzati controllati ma si applica a qualsiasi tipo di valutazione che ha lo scopo di studiare l'impatto di un intervento che viene introdotto gradualmente, o che potrebbe essere modificato a seguito della valutazione.

Esempio

Dal 1970, gli esperimenti randomizzati sono stati ampiamente implementati per misurare l'elasticità dell'offerta di lavoro agli incentivi finanziari[28].

L'SSP[29] è un esperimento randomizzato su larga scala attuato in Canada dal 1992 al 1999. Il programma offriva integrazioni ai guadagni dei genitori *single* già beneficiari di misure di integrazione al reddito per tre o più anni, a condizione che abbandonassero tali misure e tornassero al lavoro entro un anno dall'introduzione nel programma. I beneficiari sono stati selezionati casualmente dagli archivi amministrativi della misura precedente (*Assistance Insurance*). L'integrazione al reddito è stata molto generosa: la combinazione di sussidio e salario era infatti quasi doppia del reddito minimo previsto per un lavoro a tempo pieno.

È importante notare che questo progetto non valuta l'impatto dell'introduzione di un sistema di reddito minimo, ma il cambiamento delle regole relative ai livelli delle prestazioni offerte ai beneficiari del reddito minimo. Si tratta di un'importante questione di *policy*, in quanto il profilo dei livelli delle prestazioni può influenzare il comportamento dei beneficiari rendendo più o meno attraenti i redditi da lavoro. Uno dei vantaggi degli esperimenti randomizzati è la possibilità di misurare l'impatto dei diversi componenti di un dato programma. Ad esempio, lo studio sul SSP misurava l'impatto del solo incentivo finanziario, mentre lo studio sul SSP Plus misurava gli effetti degli incentivi finanziari e dei servizi di supporto alla ricerca di un'occupazione.

I ricercatori hanno scoperto che, mentre gli incentivi finanziari da soli hanno avuto un impatto positivo sulla partecipazione al mercato del lavoro e sui redditi da lavoro, la combinazione dell'integrazione dei guadagni con i servizi di consulenza per la ricerca di lavoro ha avuto effetti ancora più grandi.

Un esperimento randomizzato simile è stato implementato in Francia. I ricercatori[30] hanno verificato se un reddito minimo garantito esteso ai giovani (sotto i 25 anni), il Revenue Contractuel d'Autonomie, migliora la partecipazione ad un programma d'inserimento lavorativo aiutando i giovani a ottenere posizioni meglio pagate e più stabili. A questo scopo sono stati randomizzati alcuni beneficiari del programma[31] di attivazione ed è stato loro offerto il reddito minimo garantito. Rispetto ai soggetti inseriti nel gruppo di controllo, i beneficiari hanno diminuito

28 In Meyer (1995) si trova una panoramica delle principali lezioni apprese dall'U.S. Unemployment Insurance Experiments: http://economics.sas.upenn.edu/~hfang/teaching/socialinsurance/readings/fudan_hsbc/Meyer95(4.13).pdf

29 http://www.srdc.org/what-we-do/demonstration-projects-impact-evaluation-studies/self-sufficiency-project.aspx

30 I ricercatori sono Romain Aeberhardt, Véra Chiodi, Bruno Crépon, Mathilde Gaini e Augustin Vicard.

31 Partire da individui già inclusi in un programma è un modo per evitare qualsiasi impatto sul flusso e concentrarsi sull'impatto del programma su individui identici.

la propria offerta di lavoro, ma solo nei primi mesi successivi all'introduzione nel sistema. Inoltre, i beneficiari hanno partecipato più regolarmente al programma di attivazione, hanno visto aumentare il proprio reddito disponibile ma non hanno modificato lo stato occupazionale, dopo tre mesi.

Applicabilità alla riforma ipotizzata

Programmando esperimenti randomizzati, anziché basarsi su analisi di dati osservazionali, i ricercatori e i decisori politici non hanno bisogno di adattare le loro domande ai dati già esistenti. Al contrario, sono liberi di concentrarsi sulle questioni più rilevanti e di progettare la strategia di raccolta dei dati che è più adatta a rispondere alle domande di valutazione. Ad esempio, un progetto sperimentale nel quale gli individui vengono suddivisi a caso in diversi gruppi e a ciascun gruppo viene offerto un differente elemento del programma può aiutare a comprendere l'impatto di ciascuna delle specifiche misure di cui si compone la politica. A questo scopo è possibile confrontare persone in luoghi diversi (selezionando a caso distretti che offrono varianti differenti) o persone nello stesso luogo (selezionando a caso individui all'interno dei distretti).

In questa riforma, un possibile candidato per una valutazione di questo tipo sono le politiche attive del lavoro. Un certo numero di studi randomizzati sono stati eseguiti in diversi paesi per valutare l'impatto degli interventi senza sollevare particolari difficoltà. Una strategia naturale è assegnare in modo casuale la possibilità di accedere al programma di attivazione, considerando gli esclusi come gruppo di controllo. La fattibilità di tale sistema dovrebbe essere controllata rispetto a leggi e regolamenti.

È anche possibile testare l'impatto di diverse varianti dei benefici (o delle modalità con per accedere a prestazioni supplementari). Ciò può essere fatto assegnando casualmente differenti versioni del programma a una serie di aree/distretti comprese nell'esperimento. Per immaginare un disegno simile sarà necessario un numero di aree/distretti sufficientemente grande.

Riferimenti bibliografici

Blundell R., Brewer M., Shephard A. (2005), *Evaluating the Labour Market Impact of Working Families' Tax Credit using difference-in-differences*. London: HM Revenue and Customs.

Bourguignon F. (2009), *Rapport final sur l'évaluation des expérimentations rSa - Comité d'Evaluation des experimentations*. Paris: Haut commissaire pour la solidarité active contre la pauvreté.

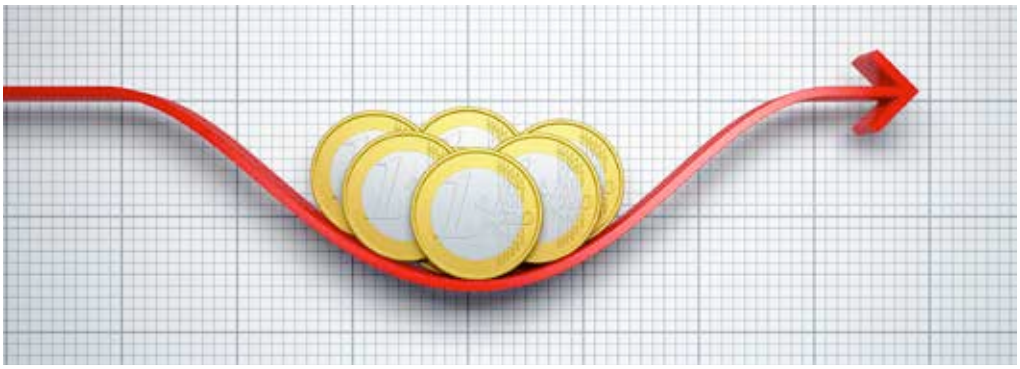
Eissa N., Liebman J.B. (1996), Labor Supply Response to the Earned Income Tax Credit. *The Quarterly Journal of Economics*, 112, 2: 605-637.

Jones M.R. (2013), *The EITC and Labor Supply: Evidence from a Regression Kink Design*. Washington: Center for Administrative Records Research and Applications, U.S. Census Bureau.

Michalopoulos C., Robins P.K., Card D. (2005), When financial work incentives pay for themselves: evidence from a randomized social experiment for welfare recipients. *Journal of Public Economics*, 89: 5-29.

Michalopoulos C., Tattrie D., Miller C., Robins P.K., Morris P., Gyarmati D., Redcross C., Foley K., Ford R. (2002), *Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients*. Ottawa: Social Research Demonstration Corporation.

Ying L., Michalopoulos C. (2001), *SSP Plus at 36 Months: Effects of Adding Employment Services to Financial Work Incentives*. Ottawa: Social Research Demonstration Corporation.



ESEMPIO 3

ESEMPIO 3 - COME VALUTARE UNA RIFORMA DELL'ASSISTENZA A LUNGO TERMINE?

1. Introduzione

Questo caso di studio mostra come valutare l'impatto di un programma di assistenza a lungo termine (*Long term care o LTC*). Sono quindi presentati i principali disegni di ricerca disponibili e il modo come questi sono stati utilizzati in passato. Il caso studio mostra che questi disegni variano essenzialmente per: 1. i loro requisiti metodologici e 2. Le ipotesi che devono essere fatte per quanto riguarda la comparabilità dei beneficiari e dei gruppi di controllo. Questa discussione è illustrata da esempi condotti negli Stati Uniti, nel Regno Unito e in Slovenia. Il caso di studio è organizzato come segue: il paragrafo 2 fornisce una panoramica delle politiche sanitarie a lungo termine; il paragrafo 3 descrive brevemente le caratteristiche di una riforma tipo; il paragrafo 4 spiega come costruire situazioni controfattuali, e il paragrafo 5 discute dei diversi metodi che potrebbero essere applicati, illustrati con esempi concreti.

2. La gestione della cura a lungo termine, una panoramica

Nel 2050 si prevede il raddoppio del numero degli anziani di età superiore agli 80 anni. La quota relativa sulla popolazione aumenterà dal 4,7% all'11,3% nei 27 Stati membri dell'UE. Tra un quarto e la metà di loro avrà bisogno di aiuto nella vita quotidiana (OCSE / Commissione europea, 2013). I sistemi sanitari sono spesso mal equipaggiati per rispondere al rapido aumento dei pazienti con problemi di salute multipli, tra cui la riduzione delle capacità funzionali e cognitive. La cura per queste persone potrebbe frammentarsi tra diversi professionisti e organizzazioni, con conseguenti rischi per la qualità e la sicurezza derivanti da duplicazioni o omissioni di cura. Ciò ha portato a inviti diffusi a favore di una maggiore integrazione delle cure (Curry e Ham, 2010).

Il *case management* è una funzione fondamentale di integrazione ed è sempre combinato con l'uso di strumenti per identificare i pazienti a rischio di eventi avversi (Lewis, Curry & Bardsley, 2011). Il *case management* è definito come un "approccio proattivo alla cura che comprende l'accertamento, la valutazione, la pianificazione e il coordinamento della cura" (Ross, Curry & Goodwin, 2011).

Le evidenze degli effetti del *case management* sono 'promettenti, ma non univoche' (Purdy, 2010). Ciò soprattutto a causa della difficoltà di attribuire impatti tangibili (ad esempio la riduzione nell'utilizzo dei ricoveri ospedalieri) agli interventi di *case management* quando più fattori sono in gioco. Questo problema di attribuzione è comune nella valutazione dei programmi per ridurre l'ospedalizzazione (Steventon *et al*., 2011; Purdy, 2010). Un'ulteriore complicazione nella valutazione dell'impatto del *case management* è che esso non rappresenta un intervento standardizzato; i programmi possono infatti variare notevolmente, il che rende difficile fare paragoni o trarre conclusioni generalizzate. Gli impatti del *case management* possono anche essere difficili da quantificare (per esempio, l'impatto sull'esperienza vissuta dal paziente e sulla sua salute). Inoltre, l'impatto potrebbe non essere misurabile nel breve termine, aumentando le difficoltà di attribuzione dell'effetto.

Vi è, tuttavia, evidenza che un *case management* opportunamente progettato e realizzato può avere un impatto positivo su:

- le esperienze dei pazienti;
- la salute dei pazienti, tra cui la qualità della vita, l'indipendenza, la funzionalità e il benessere generale;
- l'utilizzazione dei servizi, tra i quali i ricoveri ospedalieri, la durata del ricovero e i ricoveri per l'assistenza a lungo termine (si veda la rassegna in: Ross, Curry e Goodwin, 2011).

Il *case management* si è rivelato particolarmente efficace quando inserito in un programma più ampio in cui l'impatto cumulato di molteplici strategie (non di un singolo intervento) può avere successo nel migliorare le esperienze di cura e la salute dei pazienti (Powell-Davies *et al*., 2008; Ham, 2009). Nonostante l'evidenza controversa, è ampiamente riconosciuto che il *case management* sia un valido approccio per la gestione di persone con esigenze molto complesse e di lungo termine ed è quindi ampiamente utilizzato in questi casi (LTC).

3. La riforma

Una riforma in linea con i principi ricordati sopra avrebbe l'obiettivo di ridurre le disuguaglianze e la frammentazione dell'assistenza a lungo termine attraverso lo sviluppo di servizi domiciliari e di comunità, unificando quelli di natura sanitaria e assistenziale. Cambiamenti più specifici dovrebbero includere:

- un punto unico di ingresso per i pazienti;
- una procedura uniforme per la valutazione dei bisogni;
- un processo per la preparazione di piani di cura individuali; e
- la formazione per gli assistenti informali.

La persona che ha bisogno di cure a lungo termine dovrebbe poi decidere se optare per servizi in natura o trasferimenti in denaro. La soglia, la portata, il contenuto di diritti e prestazioni sono elementi molto importanti.

4. Come costruire controfattuali

Al fine di valutare l'impatto del *case management*, è necessario costruire un controfattuale[32]; cioè confrontare i beneficiari del nuovo regime con i beneficiari simili di misure pre-esistenti (non integrate). Oltre a consentire il confronto del nuovo schema con quelli vecchi, i controfattuali possono essere utilizzati anche per valutare gli effetti dei diversi aspetti della stessa riforma, in modo da individuare i più efficienti. Ciò è molto importante da un punto di vista della politica, perché può consentire un'ottimizzazione della spesa concentrandola sulle opzioni con il maggiore impatto sui risultati desiderati. Inoltre, i controfattuali aiutano a stimare l'eterogeneità dell'impatto sulle diverse sottopopolazioni (ad esempio, uomini e donne), consentendo di costruire azioni su misura.

5. Potenziale dei diversi metodi di valutazione d'impatto controfattuale

5.1. Abbinamento statistico

Descrizione

Con questo metodo si costruisce un controfattuale abbinando partecipanti e non partecipanti sulla base di una serie di caratteristiche osservabili. Un abbinamento di successo richiede una ricerca preliminare in grado di identificare le variabili che potrebbero essere statisticamente correlate alla probabilità di partecipare al programma e ai relativi risultati. Per creare corrispondenze sufficienti sono necessari campioni consistenti. Il metodo fornisce una stima dell'effetto di un intervento per tutti i partecipanti che possono essere correttamente abbinati a un non partecipante.

Assunzioni

L'assunzione principale di questo metodo è che tutte le caratteristiche di base che influenzano partecipazione e risultati d'interesse possano essere osservate e valutate.

Esempio

Challis e Davies (1985) hanno usato questo metodo per valutare l'impatto del Community Care Scheme nel Kent (UK). Il programma ha tentato di affrontare il problema della scarsa qualità dei servizi e della loro frammentazione utilizzando due strategie separate ma correlate: (i) una maggiore flessibilità di risposta al bisogno, in modo da migliorare il contesto in cui il servizio è erogato; e (ii) una migliore gestione dei casi, attraverso l'assegnazione di una chiara responsabilità a un unico soggetto che abbia la missione di integrare i differenti servizi in un 'pacchetto di cura' coerente.

Al fine di fornire una base comparativa per la valutazione, gli effetti delle cure su coloro che ricevono il nuovo programma sono stati confrontati con casi simili in aree adiacenti. I singoli casi sono stati abbinati sulla base di fattori che possono

32 Si veda a proposito; ESF Guide on Counterfactual Evaluation: Design and Commissioning of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2012..

essere predittori della sopravvivenza. Si tratta di età, sesso, composizione del nucleo familiare, presenza di stato confusionale, disabilità fisica e ricettività all'aiuto. Con questo processo sono state quindi selezionate 74 coppie, nelle quali un soggetto ha ricevuto il nuovo servizio e l'altro quello standard, sulle quali è stato svolto un confronto. La valutazione ha dimostrato che ci sono stati miglioramenti significativi sia nel benessere soggettivo sia nella qualità delle cure ricevute dai destinatari dei nuovi servizi rispetto ai beneficiari dei servizi standard.

Applicabilità alla riforma ipotizzata

Al fine di valutare l'impatto della riforma, si potrebbero abbinare beneficiari del vecchio e del nuovo programma in base a un insieme di caratteristiche osservate, come è stato fatto per il Community Care in Kent. Un uso alternativo del metodo dell'abbinamento statistico potrebbe prevedere il confronto degli effetti di varie forme di *case management* e di servizi sulle persone che beneficiano del nuovo programma, evitando di confrontare nuovo e vecchio programma.

Si noti tuttavia che, in tale contesto, l'abbinamento presenta un limite importante. Il metodo richiede, infatti, che tutte le caratteristiche rilevanti per determinare la partecipazione e influenzare i risultati di interesse possano essere osservate e rappresentate. Si tratta di un'ipotesi forte in generale e, in questo particolare caso, altamente improbabile. E', infatti, importante notare che la partecipazione al programma è volontaria, pertanto risulta difficile determinare quanta parte dell'eventuale effetto sia determinata dal sistema di servizi di assistenza e quanta dipenda da differenze pre-esistenti nelle caratteristiche osservabili e non osservabili dei partecipanti.

5.2. Confronto attorno al punto di discontinuità (RDD)

Descrizione

Questo metodo confronta individui posizionati appena al di sopra di una determinata soglia di ammissibilità continua, con quelli posizionati appena sotto. Si tratta probabilmente di individui molto simili e la soglia determina se siano o meno esposti all'intervento che si intende valutare. L'ampiezza dell'intervallo attorno alla soglia determina la dimensione del campione.

Assunzioni

Questo metodo si basa sul presupposto che l'intervento si basi su un criterio di selezione chiaramente quantificabile basato su un punteggio continuo e che i partecipanti non possono influenzare i punteggi attorno alla soglia. Inoltre, si assume che gli individui appena sotto e appena sopra la soglia non siano significativamente diversi.

Applicabilità alla riforma ipotizzata

Poiché la variabile sulla quale si applica la soglia deve essere continua, le possibilità di scelta sono piuttosto limitate. Un'opzione potrebbe consistere nella data di avvio del programma. I candidati che possono beneficiare della misura di assistenza integrata, dopo una certa data saranno assegnati al nuovo programma di *case management*, mentre quelli arrivati prima riceverebbero la versione precedente del programma. La data di soglia deve essere determinata retroattivamente, in modo che i beneficiari non possano modificare la propria scelta. Inoltre, i due programmi dovranno funzionare

parallelamente (almeno durante la fase di valutazione), in modo che i beneficiari di ciascun programma possano essere osservati durante il medesimo periodo.

Un altro approccio richiederebbe l'utilizzo di una delle variabili che determina l'ammissibilità al programma. In linea di principio, si tratta di un'opzione valida quando l'iscrizione al servizio si basa su modelli di rischio predittivi. Tali modelli utilizzano algoritmi statistici per prevedere il livello del futuro rischio di ricovero ospedaliero di un individuo (Billings *et al*., 2006; Nuffield Trust, 2011). In pratica, però, la maggior parte dei programmi utilizza una combinazione di modello predittivo e di giudizio clinico: il modello è utilizzato per identificare individui ad alto rischio e il medico esprime comunque un giudizio sull'opportunità che la persona benefici del *case management*. Anche in questo caso, si può utilizzare la soglia dell'indice di rischio per separare due popolazioni simili vicine alla soglia, anche se questa non determina completamente l'ammissibilità dell'intervento (questo disegno è chiamato *'fuzzy'* (sfocato) ed è descritto nella guida FSE sulla valutazione controfattuale (op. cit.)).

È importante sottolineare che la valutazione del rischio richiede dati di buona qualità. I più potenti modelli predittivi richiedono l'accesso ai *record* di un individuo prima del suo ricovero ospedaliero, a basi dati epidemiologiche generali (General Practice Records) e agli accessi al pronto soccorso. Anche i dati sulle prestazioni di assistenza sociale potrebbero aggiungere potere predittivo, anche se non sempre sono disponibili. Per ottenere una stima non distorta di un effetto dell'intervento è comunque necessario che il processo di assegnazione sia trasparente e perfettamente misurato (Shadish *et al.,* 2002). La premessa di base di un disegno RDD è che i partecipanti posizionati in prossimità del punto di soglia sono più simili e, quindi, costituiscono i casi migliori da confrontare per valutare l'effetto dell'intervento.

Se la soglia è rispettata rigorosa(mente, il disegno RDD tiene sotto controllo la maggior parte delle minacce alla sua validità, in quanto qualsiasi distorsione che potrebbe influenzare il gruppo di intervento dovrebbe prodursi proprio in coincidenza del valore di soglia. Sebbene ciò sia teoricamente possibile, la probabilità che tale evento si verifichi è piuttosto remota.

5.3. Differenza nelle differenze (DID)

Descrizione

Il metodo confronta nel tempo la variazione dei risultati ottenuti prima e dopo l'inizio del programma da un gruppo di partecipanti e non partecipanti. Fornisce una misura dell'impatto sull'intera popolazione dei partecipanti, controllando per condizioni costanti (osservate o meno) che possono essere correlate sia con i risultati finali sia con il fatto di essere parte del gruppo di controllo.

Con la Differenza nelle differenze è possibile sia confrontare un gruppo di persone che ha i requisiti per ricevere l'intervento con un gruppo simile ma non ammissibile, sia confrontare aree pilota nelle quali viene introdotto il programma con aree di confronto che non lo ricevono.

Assunzioni

L'approccio è basato sull'assunzione delle "dinamiche parallele". Al fine di determinare se la differenza nei risultati è dovuta al programma, si deve infatti assumere che le dinamiche dei risultati dei partecipanti e dei non partecipanti rimarrebbero uguali in assenza del programma e che, in tal caso, rimarrebbe invariata anche la composizione di ciascun gruppo. Un modo per validare l'ipotesi è verificare se entrambi i gruppi hanno mostrato andamenti paralleli prima dell'introduzione del programma. Altre modalità di verifica prevedono, da un lato l'esecuzione di test che simulano trattamenti "placebo" (assegnati casualmente a un sottoinsieme delle unità dei gruppi di intervento e controllo) oppure stimano gli effetti su risultati che non dovrebbero essere ragionevolmente attribuibili all'intervento, dall'altro utilizzano gruppi di controllo differenti.

Esempio

Nel 2008, il Dipartimento della Salute inglese ha invitato organizzazioni sanitarie che offrono approcci innovativi a candidarsi per fornire servizi integrati di cura, in seguito alle preoccupazioni, espresse in particolare dalle persone anziane, circa la crescente frammentazione delle cure. Volutamente, il governo non ha dato indicazioni sulle modalità per raggiungere tale integrazione, ma ha invece richiesto di proporre una serie di approcci diversi sviluppati 'dal basso' da parte dei fornitori di servizi assistenziali. Questo approccio ha prodotto una vasta gamma di interventi differenti, un modello comune ha però riguardato la gestione dei casi di persone anziane a rischio di ricovero ospedaliero d'urgenza. In questi interventi, le principali attività d'integrazione hanno coinvolto gli ambulatori e gli altri servizi sanitari di comunità.

Una valutazione svolta nel 2012 ha riportato i risultati dei sei siti dove sono state condotte le esperienze di *case management*; i dati riguardano le modifiche del lavoro di cura svolto dagli operatori, i cambiamenti riscontrati nell'esperienza dei pazienti e

quelli relativi ai ricoveri e ai costi (Roland *et al.,* 2012). Mediante un'analisi basata sul metodo della Differenza nelle differenze sono stati messi a confronto i dati sull'ospedalizzazione di due gruppi di pazienti per un periodo che comprendeva i sei mesi precedenti all'intervento e i sei mesi successivi. I due gruppi di pazienti differivano per il tipo d'intervento ricevuto, nuovo o vecchio. Il confronto dei dati dei due gruppi, ha mostrato che nel gruppo sperimentale si sono verificati sia un significativo aumento dei ricoveri di emergenza sia una significativa riduzione sia dei ricoveri volontari e delle visite ambulatoriali rispetto al gruppo di controllo.

Una preoccupazione riguardo a questo tipo di studi è la possibilità che esistano differenze sistematiche - non osservabili e che perciò non possono essere tenute in considerazione - tra i gruppi di intervento e di controllo. I valutatori suggerirono infatti che i due gruppi non fossero strettamente comparabili. Altri problemi rendevano inoltre difficile trarre conclusioni generali da questo studio. Per esempio, i progetti pilota costituivano un gruppo alquanto eterogeneo d'interventi che, a loro volta, sono stati ulteriormente adattati e modificati in corso d'opera, conformandosi al mutevole ambiente sanitario nel quale operavano. Così, l'idea che la valutazione abbia riguardato un singolo intervento risulta piuttosto inverosimile.

Applicabilità alla riforma ipotizzata

Il metodo della Differenza nelle differenze aiuta a controllare i fattori ambientali (nuove politiche, congiuntura economica) che potrebbero indurre un cambiamento nel gruppo sperimentale. Tuttavia, il metodo si basa sul presupposto che non vi siano altri shock-sconosciuti, oltre all'intervento, che potrebbero aver influito in modo diverso sui risultati del gruppo sperimentale e di quello di controllo. Questa è un'ipotesi forte, in quanto è molto difficile escludere l'esistenza di tali shock. Un modo per applicare il metodo alla riforma LTC sarebbe di sperimentare l'intervento in aree pilota ragionevolmente rappresentative del territorio nel suo complesso - o almeno non differenti dal punto di vista delle grandezze socio-economiche e demografiche essenziali. Le aree di controllo andrebbero poi identificate nell'ambito dello stesso gruppo di aree disponibili. In tal caso le differenze nei risultati riscontrate tra le aree pilota e di controllo dopo l'attuazione della politica sarebbe interpretabile come l'effetto dell'intervento.

L'ipotesi delle "dinamiche parallele" è difficile da giustificare se le aree pilota e di controllo sono differenti, ad esempio, perché le aree pilota dispongono di servizi sociali migliori. Questa ipotesi può essere verificata selezionando due gruppi di aree pilota entrate nella sperimentazione in tempi diversi. Se può essere dimostrato che entrambe le aree, d'intervento e di controllo, avevano ospedalizzazione o mortalità analoghe prima dell'introduzione del programma, allora si può supporre che i controlli e i beneficiari siano comparabili. Solo in questo caso qualsiasi differenza significativa riscontrata tra i due gruppi dopo l'introduzione del programma potrebbe essere considerata come un impatto dell'intervento.

5.4. Studi controllati randomizzati (RCT)

Descrizione

Gli studi controllati randomizzati misurano l'impatto medio di una politica o di un programma assegnando casualmente i soggetti ai gruppi sperimentali e di controllo e confrontando la differenza nei risultati ottenuti.

Assunzioni

In questo caso non è necessario fare affidamento su ipotesi forti come richiesto da altri protocolli, perché l'assegnazione casuale da campioni sufficientemente grandi garantisce che gli individui siano simili, in media, sia per quanto riguarda le caratteristiche osservabili sia per quelle non osservabili. L'unica assunzione necessaria è che le persone non si comporteranno in modo diverso perché consapevoli di essere inserite in un esperimento. Questa ipotesi si applica comunque a tutti i tipi di esperimento, inclusi gli studi controllati randomizzati e gli esperimenti naturali[33].

È anche importante che il sistema in corso di valutazione sia ben definito e maturo e, quindi, simile a quello che potrà essere generalizzato su scala più ampia. Anche questo aspetto non riguarda esclusivamente gli studi randomizzati controllati e si applica a qualsiasi tipo di valutazione che ha lo scopo di valutare l'impatto di un intervento che viene introdotto gradualmente, o che potrebbe essere modificato a seguito della valutazione.

Esempio

La rassegna sistematica condotta da You *et al.* (2012) ha permesso di identificare 10 studi controllati randomizzati sul tema in discussione. Uno di questi riguarda l'integrazione di servizi di cura per pazienti acuti e a lungo termine (Applebaum *et al.*, 2002). L'intervento si è basato su risorse umane dedicate, una migliore comunicazione e l'interesse del fornitore a fornire il miglior servizio possibile.

33 Gli effetti indotti dalla valutazione (o dall'osservazione) si verificano quando i soggetti modificano il proprio comportamento perché sono consapevoli di partecipare a uno studio e non a causa dell'intervento stesso.

Lo studio ha reclutato anziani disabili cronici assistiti a domicilio che correvano il rischio di utilizzare una quantità elevata di servizi per acuti. La metà dei pazienti è stata assegnata in modo casuale a un'infermiera clinica con la funzione di *care manager* che, in collaborazione con i gestori del programma, aveva il compito di migliorare il collegamento tra i servizi per acuti e di assistenza a lungo termine utilizzati dai pazienti inseriti nel programma. Un geriatra supervisionava l'attività del *care manager*.

Sebbene si fosse rilevata qualche variazione nell'intensità del ricorso alle cure sanitarie e nei costi sostenuti tra gruppo sperimentale e di controllo durante i 18 mesi dell'intervento, gli autori hanno concluso che non vi fossero differenze tra i gruppi in nessuna delle variabili di *outcome* esaminate.

Gli sforzi per integrare i sistemi di assistenza per pazienti acuti e a lungo termine si sono dimostrati più difficili del previsto. L'intervento, che ha tentato di creare integrazione attraverso un intenso servizio di *care manager*, ma senza incentivi finanziari o regolamentari, semplicemente non era abbastanza intenso da produrre un cambiamento significativo per i clienti serviti. Il programma è stato inoltre influenzato da vari cambiamenti organizzativi, come ad esempio le variazioni nella gestione degli ospedali coinvolti negli studi, che hanno ripercussioni sul loro modo di comunicare con i *care manager*.

Applicabilità alla riforma ipotizzata

L'esempio precedente è molto rappresentativo. Una caratteristica interessante di questo approccio è che i differenti impatti delle varianti della politica possono essere valutati separatamente. Ad esempio, si potrebbe immaginare un intervento con caratteristiche differenti da valutare separatamente: ad esempio confrontare l'effetto di *case manager* con piccoli gruppi di utenti (30 pazienti) o gruppi più grandi (100 pazienti) oppure l'effetto dell'erogazione di un servizio reale di cura messo a confronto con trasferimenti monetari (si veda in proposito Yordi *et al*., 1997).

Riferimenti bibliografici

Applebaum R., Straker J., Mehdizadeh S., Warshwa G., Gothelf E. (2002), Using high-intensity care management to integrate acute and long-term care services: substitute for large scale system reform? *Care Management Journal*, 3, 3: 113–119. (www.ncbi.nlm.nih.gov).

Challis D., Davies B. (1985), Long Term Care for the Elderly: the Community Care Scheme. *British Journal of Social Work*, 15, 6: 563-579. (Abstract in http://bjsw.oxfordjournals.org).

Curry N., Ham C. (2010), *Clinical and Service Integration: The route to improved outcomes*. London: The King's Fund. (www.kingsfund.org.uk).

Ham C. (2009), The ten characteristics of a high-performing chronic care system. *Health Economics, Policy and Law*, 5, 1: 71–90. (http://www.ncbi.nlm.nih.gov).

Lewis G., Curry N., Bardsley M. (2011), *Choosing a predictive risk model: a guide for commissioners in England*. London: Nuffield Trust. (www.nuffieldtrust.org.uk).

OECD, European Commission (2013), *A Good Life in Old Age? Monitoring and Improving Quality in Long-term Care*. Paris: OECD Publishing. (www.oecd.org).

Powell-Davies G., Williams A., Larsen K., Perkins D., Roland M., Harris M. (2008), Coordinating primary health care: an analysis of the outcomes of a systematic review. *Medical Journal of Australia*, 188, 8: S65–S68. (www.ncbi.nlm.nih.gov).

Purdy S. (2010), *Avoiding Hospital Admissions: What does the research say?* London: The King's Fund. (www.kingsfund.org.uk).

Roland M. (ed.) (2012), Case management for at-risk elderly patients in the English integrated care pilots: observational study of staff and patient experience and secondary care utilisation. *International Journal of Integrated Care*, July, 12. (www.ijic.org/index.php).

Ross S., Curry N., Goodwin N. (2011), *Case management. What is it and how can it best be implemented?* London: The King's Fund. (www.kingsfund.org.uk).

Steventon A., Bardsley M., Billings J., Georghiou T., Lewis G.H. (2011), *A Case Study of Eight Partnership for Older People Projects (POPP): an evaluation of the impact of community-based interventions on hospital use*. London: Nuffield Trust. (www.nuffieldtrust.org.uk).

Yordi C., DuNah R., Bostrom A., Fox P., Wilkinson A., Newcomer R. (1997), Caregiver Supports: Outcomes from the Medicare Alzheimer's Disease Demonstration. *Health Care Financial Review*, 19, 2: 97-116. (www.ncbi.nlm.nih.gov).

You E.C. (ed.) (2012), Effects of case management in community aged care on client and career outcomes: a systematic review of randomized trials and comparative observational studies. BMC Health Services Research, 12: 395. (www.biomedcentral.com).

DEFINIZIONI

Assunto	Relazioni causa-effetto accettate come valide, o stime dell'esistenza di un fatto dall'esistenza nota di un altro/altri fatto/i.
Baseline	È la situazione standard rispetto alla quale si misurano tutte le successive modifiche generate da un intervento.
Campione probabilistico	Un metodo di campionamento probabilistico è qualsiasi metodo di campionamento che utilizza una qualche forma di selezione casuale (Trochim, 2007).
Controfattuale	È la stima della situazione (ipotetica) in cui si sarebbero trovate le persone inserite in un programma se quel programma non fosse mai stato attuato. Questa nozione è utilizzata per comprendere l'effetto causale del programma (Glennerster, Takavarasha 2013).
Dimensione dell'effetto	È una misura che descrive la grandezza della differenza rilevata tra il gruppo sperimentale e quello di controllo nelle variabili risultato.
End-line	È la misura delle variabili risultato rilevata alla fine dello studio.
Esperimento	Un esperimento è una procedura effettuata con lo scopo di verificare, confutare, o stabilire la validità di un'ipotesi. Gli esperimenti forniscono informazioni sulle relazioni causa-effetto dimostrando quali risultati si realizzano quando un particolare fattore viene modificato.
Input	È una risorsa o un fattore di produzione (lavoro, capitale) utilizzato nella produzione (output) di un'organizzazione.
Intervento (politica)	Misure adottate per migliorare un problema sociale.
Meta-valutazione (o meta-analisi)	È l'uso di metodi statistici per combinare i risultati di singoli studi (Cochrane Collaboration).
Programma	In una politica pubblica, un programma fa riferimento a una serie di interventi combinati tra loro.

DEFINIZIONI

Protocollo	Un protocollo è il piano particolareggiato di uno studio. Per convenzione, è scritto secondo il seguente formato: i) Titolo del progetto; ii) Sintesi del progetto; iii) Descrizione del progetto (motivazione, obiettivi, gestione e analisi dei dati; metodologia); iv) Considerazioni etiche; v) Riferimenti.
Rassegna sistematica	Una rassegna sistematica cerca di identificare, valutare e sintetizzare tutta l'evidenza empirica che rispetta i criteri di ammissibilità pre specificati per rispondere a una determinata domanda di ricerca. I ricercatori che effettuano rassegne sistematiche utilizzano metodi chiari volti a minimizzare le distorsioni, al fine di produrre i risultati più affidabili da utilizzare per supportare il processo decisionale (Cochrane Handbook for Systematic Reviews of Interventions).
Teoria del cambiamento (Theory of Change - TOC)	È una metodologia specifica per la pianificazione, la partecipazione e la valutazione utilizzata nel settore filantropico e nella pubblica amministrazione per promuovere il cambiamento sociale. La teoria del cambiamento definisce obiettivi a lungo termine e mappe retrospettive per identificare le condizioni necessarie (Brest, 2010).
Triangolazione	Nelle scienze sociali, il termine triangolazione è spesso usato per indicare che in uno studio sono utilizzati due (o più) metodi per verificarne i risultati. L'idea è che si può essere più sicuri della validità del lavoro se metodi diversi portano allo stesso risultato.
Validità delle conclusioni	È il grado in affidabilità delle conclusioni riguardo alle relazioni tra i dati (Trochim, Donnelly, 2007).
Validità di costrutto	La validità di costrutto si riferisce alla misura in cui è possibile legittimamente rilevare delle inferenze dall'operazionalizzazione in uno studio da i suoi costrutti teorici (Trochim, Donnelly, 2007).
Validità esterna	È il grado in cui le conclusioni di uno studio sono estendibili ad altre persone in altri luoghi e in altri momenti (Trochim, Donnelly, 2007).
Validità interna	È una proprietà degli studi scientifici che riflette la misura in cui è garantita la conclusione causale che ne deriva.

BIBLIOGRAFIA

Adam S., Emmerson C., Frayne C., Goodman A. (2006), Early quantitative evidence on the impact of the Pathways to Work pilots. Institute for Fiscal Studies and Department for Work and Pensions, Research Report 354. Norwich: Stationery Office.

Applebaum R., Straker J., Mehdizadeh S., Warshwa G., Gothelf E. (2002), Using high-intensity care management to integrate acute and long-term care services: substitute for large scale system reform? *Care Management Journal*, 3, 3: 113–119. (www.ncbi.nlm.nih.gov).

Brest P. (2010), The Power of Theories of Change. *Stanford Social Innovation Review*, Spring.

Gertler P.J., Martinez S., Premand P., Rawlings L.B., Vermeersch C.M.J. (2011), *Impact Evaluation in Practice*. Washington: The World Bank. (http://www.worldbank.org/).

Glennerster R., Takavarasha K. (2013), *Running randomized evaluations: A practical guide*. Princeton, NJ: Princeton University Press.

Haynes L., Service O., Goldacre B., Torgerson D. (2012), *Test, Learn, Adapt. Developing Public Policy with Randomised Controlled Trials*. London: Cabinet Office, Behavioural Insights Team. (www.gov.uk).

HM Treasury (2011), *The Magenta Book – Guidance for evaluation*. (www.gov.uk).

J-PAL Europe (2011), *Social Experimentation: A methodological guide for policy makers*. (http://ec.europa.eu).

Jalan J., Ravallion M. (2003), Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. *Journal of Business & Economic Statistics*, 21, 1: 19-30. (http://amstat.tandfonline.com).

Kostøl A.R., Mogstad M. (2014), How Financial Incentives Induce Disability Insurance Recipients to Return to Work. *American Economic Review*, 104, 2: 624-655.

BIBLIOGRAFIA

Morris S. (ed.) (2004), *Designing a Demonstration Project: An Employment Retention and Advancement Demonstration for Great Britain*. London: Cabinet Office. (http://goo.gl/FyrGni).

Morris S., Tödtling-Schönhofer H., Wiseman M. (2013), *Design and Commissioning of Counterfactual Impact Evaluations - A Practical Guidance for ESF Managing Authorities*. Brussels: European Commission, DG Employment. (http://ec.europa.eu).

Torgerson D.J., Torgerson C.J. (2008), *Designing Randomised Trials in Health Education and the Social Sciences*. Basingstoke: Palgrave McMillan.

Trochim W.M., Donnelly J.P. (2007), *The Research Methods Knowledge Base, 2nd Edition*. (http://www.socialresearchmethods.net/kb/ - version current as of October 20).

Trochim W.M.K., *Research Methods Knowledge Base*. Thomson Custom Pub.

World Bank, *Impact Evaluation Toolkit*. (http://web.worldbank.org).

Yordi C., DuNah R., Bostrom A., Fox P., Wilkinson A., Newcomer R. (1997), Caregiver Supports: Outcomes From the Medicare Alzheimer's Disease Demonstration. *Health Care Financing Review*, 19, 2: 97-117.

You E.C. (ed.) (2012), Effects of case management in community aged care on client and career outcomes: a systematic review of randomized trials and comparative observational studies. BMC Health Services Research, 12: 395. (www.biomedcentral.com).

I QUADERNI DELL'OSSERVATORIO

Nella Collana QUADERNI DELL'OSSERVATORIO sono stati pubblicati i seguenti titoli, scaricabili sul sito www.fondazionecariplo.it/osservatorio.

Quaderno N.1 – Periferie, cultura e inclusione sociale

Quaderno N.2 – Il valore potenziale dei lasciti alle istituzioni di beneficenza

Quaderno N.3 – Stranieri si nasce...e si rimane?

Quaderno N.4 – Oltre la famiglia: strumenti per l'autonomia dei disabili

Quaderno N.5 – L'educazione finanziaria per i giovani

Quaderno N.6 – Ricerca scientifica in ambito biomedico

Quaderno N.7 – Servizi per l'infanzia

Quaderno N.8 – Assicurazione per persone con disabilità e loro famiglie

Quaderno N.9 – Progetti e politiche per la mobilità urbana sostenibile

Quaderno N.10 – Le organizzazioni culturali di fronte alla crisi

Quaderno N.11 – I Social Impact Bond

Quaderno N.12 – Lavoro e Psiche. Un progetto sperimentale per l'integrazione lavorativa di persone con gravi disturbi psichiatrici

Quaderno N.13 – Il bando "Audit energetico degli edifici di proprietà dei comuni piccoli e medi"

Quaderno N.14 – Infrastrutture di ricerca in italia

Quaderno N.15 – Performance economica e sociale delle istituzioni di microfinanza: alcune evidenze empiriche

Quaderno N.16 – Cessione della nuda proprietà da parte di soggetti fragili: il possibile ruolo di un soggetto dedicato

Quaderno N.17 – Abitare leggero. Verso una nuova generazione di servizi per anziani

Quaderno N.18 – Progetti culturali e sviluppo urbano. Visioni, criticità e opportunità per nuove politiche nell'area metropolitana di Milano

Quaderno N.19 – Sperimentare politiche sociali innovative - Manuale introduttivo

doi: 10.4460/2015quaderno19

fondazione
cariplo