Valutazione

28

SPERIMENTA, IMPARA, ADATTA
Sviluppare politiche pubbliche con gli esperimenti randomizzati controllati

QUADERNI DELL'OSSERVATORIO

fondazione cariplo

SPERIMENTA, IMPARA, ADATTA

Sviluppare politiche pubbliche con gli esperimenti randomizzati controllati

A cura dell'Ufficio Osservatorio e valutazione della Fondazione Cariplo
Traduzione italiana di "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials"

Collana "Quaderni dell'Osservatorio" n. 28 Anno 2018

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

doi: 10.4460/2018quaderno28

INDICE

INDICE

RIQUADRI

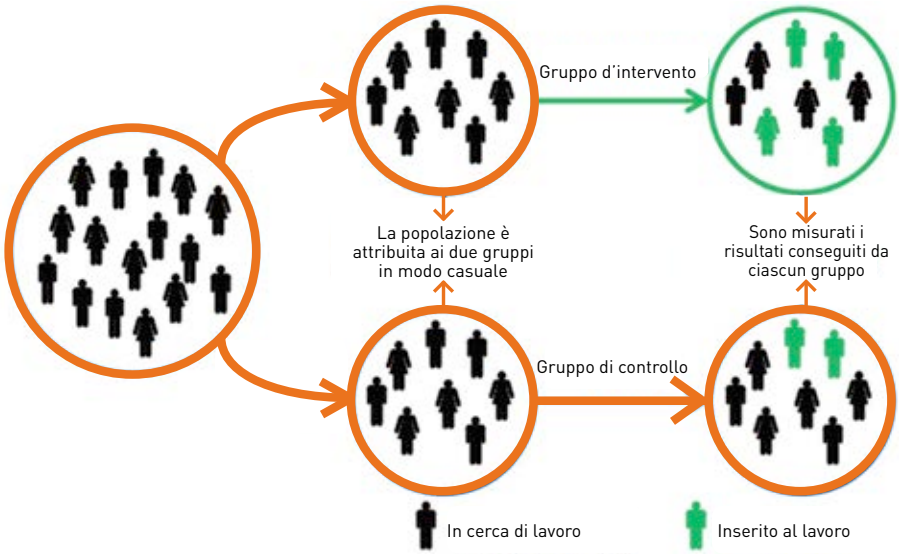
EXECUTIVE SUMMARY

Gli studi randomizzati controllati (in inglese Randomised Controlled Trials - RCT in sigla d'ora in poi) sono il modo migliore per capire se un intervento (una politica pubblica) funziona, producendo gli effetti desiderati. Sono sempre più ampiamente utilizzati nelle politiche di sviluppo, nella medicina e anche nel mondo degli affari per identificare le politiche, i farmaci o i metodi di vendita più efficaci. Questi studi costituiscono la base del metodo di lavoro del *Behavioural Insights Team*, promotore di questo lavoro. Ciononostante, gli RCT non sono ancora lo standard per valutare l'efficacia delle politiche pubbliche in Europa. Chi scrive pensa invece che dovrebbero esserlo e, con questo lavoro, cercherà di dimostrare perché e a quali condizioni.

Ciò che distingue gli studi randomizzati dagli altri metodi di valutazione è l'utilizzo di un gruppo di controllo identificato in modo casuale; ciò consente di misurare l'efficacia di un nuovo intervento mettendone a confronto i risultati con una stima di quanto sarebbe successo se non si fosse cambiato nulla. Il ricorso a un gruppo di controllo elimina tutte le distorsioni che normalmente complicano il processo di valutazione - per esempio, se si introduce un nuovo intervento per il reinserimento al lavoro dei disoccupati, come si può sapere se i beneficiari non avrebbero comunque trovato un occupazione? Nell'esempio (fittizio) presentato nella figura possiamo vedere che le persone che hanno beneficiato del nuovo intervento hanno trovato più frequentemente un lavoro rispetto ai non beneficiari. È proprio grazie al gruppo di controllo che possiamo attribuire l'effetto all'intervento di reinserimento al lavoro e non a qualche altro fattore (ad esempio, le condizioni economiche in generale miglioramento). Con un idoneo supporto da parte del mondo accademico e dei *policy maker* gli studi randomizzati possono dimostrarsi molto più economici e più semplici da realizzare di quanto si pensi. Consentendo di dimostrare quanto una politica sta funzionando, questi studi possono far risparmiare denaro e costituiscono un potente strumento per aiutare operatori e *policy maker* a identificare le politiche più efficienti o quelle che risultano meno efficaci del previsto. Questo strumento è particolarmente importante nei periodi di contrazione dei bilanci pubblici perché consente di investire nelle politiche che mostrano il migliore rapporto tra risultati prodotti e spesa sostenuta.

Abbiamo identificato le nove fasi specifiche necessarie a impostare qualsiasi studio randomizzato controllato. Molti di questi passaggi saranno già familiari a chiunque abbia provato a disegnare un progetto di valutazione rigoroso - per esempio, la necessità di chiarire fin da subito l'obiettivo specifico che la politica sta cercando di raggiun-

Disegno di uno studio randomizzato controllato per un programma di reinserimento al lavoro



gere. Alcuni passaggi - in particolare la necessità di assegnare casualmente individui o istituzioni ai diversi gruppi che ricevono interventi differenti - conferiscono agli studi randomizzati l'efficacia che viene loro riconosciuta. Queste nove fasi sono al centro della metodologia del Behavioural Insights Team "Sperimenta, impara e adatta", che si concentra sulla identificazione di ciò che funziona meglio e al suo costante miglioramento grazie all'apprendimento continuo. Queste fasi sono descritte nel riquadro 1.

Riquadro 1 – Sperimenta, impara, adatta

Sperimenta

1. Identifica due o più interventi da comparare (ad esempio vecchia e nuova politica; diverse varianti di una politica)
2. Determina esattamente il risultato (*outcome*) che la politica si propone di ottenere e come sarà misurato durante l'esperimento
3. Decidi l'unità di randomizzazione nei gruppi di intervento e di controllo: individui, istituzioni (ad esempio scuole) o aree geografiche/amministrative
4. Determina il numero di unità (persone, istituzioni o aree) necessarie per ottenere risultati statisticamente significativi
5. Assegna le unità di randomizzazione a ciascuno degli interventi utilizzando un metodo di randomizzazione robusto
6. Avvia gli interventi per i destinatari assegnati

Impara

7. Misura i risultati e valuta l'effetto degli interventi

Adatta

8. Adatta il tuo intervento tenendo conto dei risultati ottenuti
9. Ritorna al punto 1 per continuare a migliorare

INTRODUZIONE

Gli RCT sono il modo migliore per valutare se una politica funziona. Sono utilizzati da oltre 60 anni per confrontare l'efficacia di nuovi farmaci[1]. Sono inoltre sempre più impiegati nei progetti di cooperazione internazionale per analizzare, ad esempio, il rapporto costi/efficacia di diversi strumenti di lotta alla povertà (Banerjee, Duflo, 2011; Karlan, Appel, 2011). Sono ampiamente sfruttati anche dalle aziende che vogliono sapere, ad esempio, quale *layout* di un sito *web* genera più vendite. Tuttavia, gli RCT non sono ancora una pratica comune per la valutazione delle politiche pubbliche (vedi figura). Questo documento proverà a dimostrare che gli RCT si dovrebbero e potrebbero utilizzare molto di più per valutare l'efficacia di politiche vecchie e nuove o delle loro varianti, in modo da capire ciò che funziona e ciò che non funziona e per migliorare costantemente le politiche in termini di qualità e di efficacia.

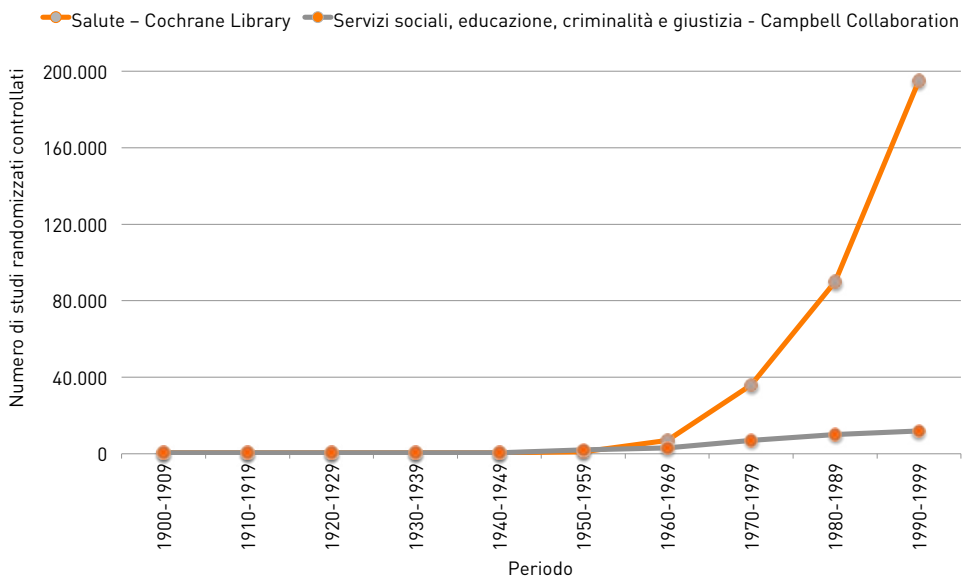
La prima parte di questo documento fornisce una definizione di studio randomizzato controllato, cercando di spiegarne l'importanza. Esso risponde a molti degli argomenti più spesso utilizzati contro l'utilizzo di questo metodo nelle politiche pubbliche e mostra che le sperimentazioni sono meno difficili da mettere in pratica di quanto normalmente si pensi e, anzi, possono costituire metodi altamente efficienti per la valutazione dei risultati delle politiche e del rapporto tra risultati e costi.

La seconda parte del documento delinea i 9 passaggi chiave necessari a condurre qualsiasi studio controllato randomizzato. Molti di questi passaggi dovrebbero essere comuni ad ogni politica, altri richiedono il supporto di esperti (accademici o meno) o di competenze specifiche all'interno della pubblica amministrazione. La filosofia "sperimenta, impara, adatta" promossa da questo documento riflette il tipo di lavoro del Behavioural Insights Team. Chi scrive crede che questo approccio possa essere utilizzato in quasi tutti i campi di azione delle politiche pubbliche:

Sperimentare un intervento significa garantire che siano state messe in atto misure robuste per valutarne l'efficacia.

1 Il primo studio controllato randomizzato pubblicato in medicina è accreditato a Sir A. Bradford Hill, un epidemiologo del Medical Research Council inglese. Il processo, pubblicato sul British Medical Journal nel 1948, ha valutato se la streptomicina è efficace nel trattamento della tubercolosi.

Studi randomizzati controllati realizzati nel 20° secolo nei settori della salute, servizi sociali, istruzione, criminalità e giustizia (Shepherd, 2007)



Imparare fa riferimento ai risultati dell'intervento: identificare 'cosa funziona' e capire se l'effetto ha una dimensione sufficientemente grande da offrire un buon rapporto costo/efficacia.

Adattare significa utilizzare gli apprendimenti per modificare l'intervento (se necessario) in modo da perfezionarne il disegno e l'implementazione sul campo.

1

1. COS'È UNO STUDIO CONTROLLATO RANDOMIZZATO?

Spesso vogliamo sapere quale, tra due o più interventi, è più efficace nel conseguire uno specifico risultato misurabile. Ad esempio, potremmo voler confrontare un nuovo intervento con quello normalmente utilizzato, o confrontare tra loro diversi livelli di "dosaggio" di un intervento (ad esempio, visite a domicilio a una madre adolescente una volta o due volte alla settimana).

Di solito, per capire se un intervento produce un beneficio, lo si implementa e si cerca di osservarne i risultati. Ad esempio, si potrebbe avviare un programma intensivo per il reinserimento al lavoro e controllare se i partecipanti trovano un'occupazione più velocemente rispetto a quanto avveniva prima dell'introduzione del programma.

Tuttavia, questo approccio soffre di una serie di inconvenienti che rendono difficile capire se il risultato osservato è effettivamente attribuibile all'intervento o è invece prodotto da qualche altro fattore esterno non controllato. Tornando all'esempio, nel caso di una forte crescita economica ci si può aspettare che le persone trovino più facilmente un impiego, indipendentemente dal nuovo intervento.

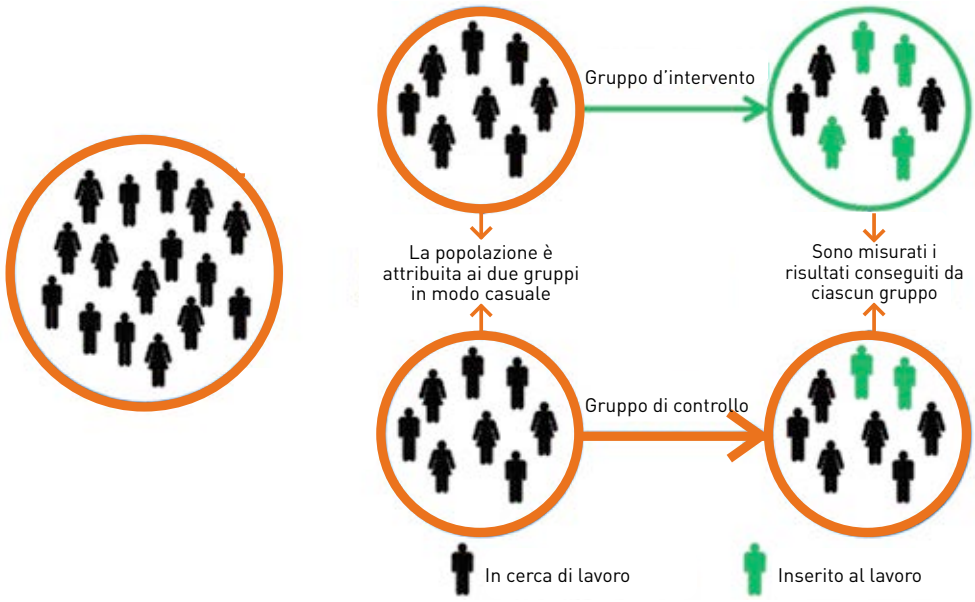
Una sfida analitica ancora più complessa ha a che fare con la cosiddetta "distorsione da selezione" (*selection bias*); le persone che decidono di partecipare a un nuovo programma d'inserimento al lavoro sono infatti sistematicamente diverse da quelle che non vogliono parteciparvi. Potrebbero, ad esempio, essere semplicemente più motivate a trovare lavoro: in questo caso i benefici del nuovo intervento risulterebbero sovrastimati. Esistono tecniche statistiche per cercare di tenere conto delle differenze di partenza tra i gruppi che ricevono diversi interventi, ma sono sempre imperfette e possono introdurre altre distorsioni.

Gli RCT risolvono questo problema facendo in modo che gli individui o i gruppi di persone che ricevono gli interventi siano, in origine, il più simili possibile. Nell'esempio del reinserimento al lavoro, si potrebbero individuare 2.000 persone ammissibili al nuovo programma e, in modo casuale, dividerli in due gruppi di 1.000 individui ciascuno: un gruppo riceverebbe l'intervento tradizionale, mentre l'altro sperimenterebbe il nuovo intervento. Assegnando a caso le persone ai due gruppi possiamo evitare che fattori esterni influenzino i risultati e dimostrare che le eventuali differenze

riscontrate tra i due gruppi sono effettivamente prodotte dalle differenze negli interventi che ricevono.

La seconda parte di questo documento descrive in dettaglio come eseguire un esperimento randomizzato controllato; al cuore di qualsiasi esperimento di questo tipo vi sono sempre alcuni specifici elementi caratterizzanti. Lo studio controllato funziona dividendo una popolazione in due o più gruppi in modo casuale (randomizzato), assegnando a ciascuno dei due gruppi un intervento differente e misurando il risultato (dichiarato anticipatamente) ottenuto da ciascun gruppo. Questo processo è riassunto nella figura 1.1.

Figura 1.1 - Studio randomizzato controllato per la valutazione di un programma di reinserimento al lavoro (esito positivo)

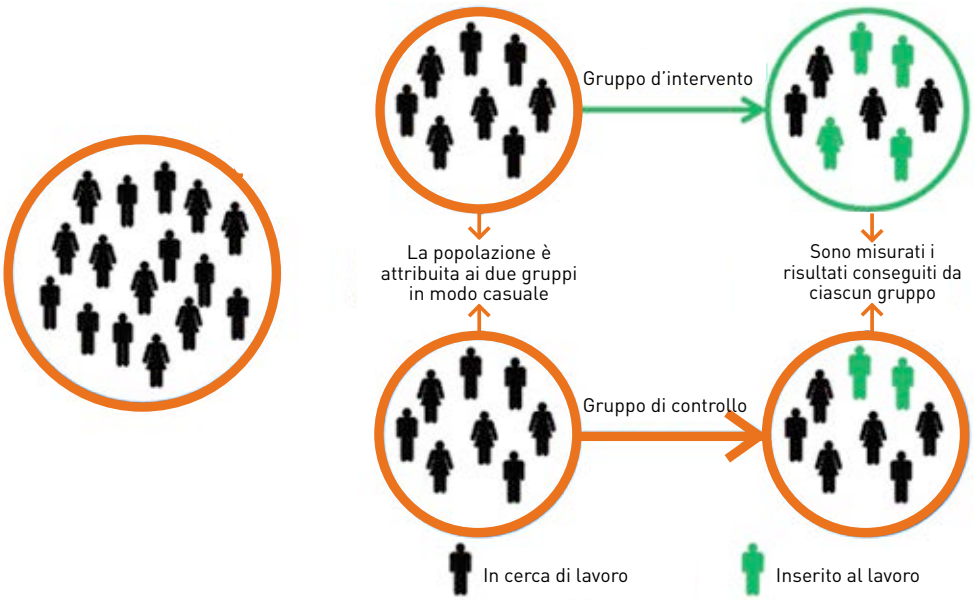


Immaginiamo di testare il nuovo programma di reinserimento lavorativo per i disoccupati. La popolazione sulla quale si svolge la valutazione è divisa casualmente in due gruppi. Solo i membri di uno dei due gruppi ricevono il nuovo programma di reinserimento ("gruppo di intervento"). Le persone dell'altro gruppo ("di controllo") ricevono invece il supporto tradizionale. In questo caso, il gruppo di controllo corrisponde a quello che riceve il placebo in uno studio clinico.

Nell'esempio riportato in figura 1.1, le persone che hanno trovato un lavoro a tempo pieno 6 mesi dopo l'inizio della sperimentazione sono indicate in verde. Lo studio mostra che le persone rientrate al lavoro sono molto più numerose nel gruppo che ha beneficiato del nuovo programma che nel gruppo di controllo. È importante notare che anche nel gruppo di controllo hanno trovato lavoro due persone, forse proprio grazie al beneficio ricevuto nell'ambito dell'intervento tradizionale.

Se il nuovo programma di inserimento al lavoro non fosse migliore del servizio fornito tradizionalmente, avremmo visto esiti simili nel gruppo di intervento e nel gruppo di controllo. La figura 1.2 mostra uno scenario di questo tipo.

Figura 1.2 - Studio randomizzato controllato per valutare un programma di reinserimento al lavoro (esito neutro)



In questo caso, i risultati dello studio dimostrano che il nuovo, costoso, programma non è migliore di quello attuale. Se non ci fosse stato il gruppo di controllo, avremmo potuto erroneamente attribuire al nuovo programma i nuovi posti di lavoro e, su tale base, decidere di metterlo a regime. Nella realtà, un errore come questo è stato evitato dal Ministero del lavoro e delle pensioni (Department for Work and Pensions - DWP) attraverso uno studio randomizzato che ha consentito di confrontare il rapporto costo-efficacia di diversi interventi.

Ogni volta che fattori esterni possano influenzare i risultati di una politica, vale sempre la pena di utilizzare un esperimento randomizzato controllato per testare l'efficacia dell'intervento prima di implementarlo su tutta la popolazione. In caso contrario, è facile scambiare ciò che avrebbe potuto verificarsi comunque per l'impatto dell'intervento.

Il nostro esempio del programma di reinserimento al lavoro cerca di capire quale tra i due interventi promossi su larga scala sia più efficace. In molti casi, uno studio controllato randomizzato non è rivolto necessariamente al problema principale cui è indirizzata la politica. Lo stesso metodo può, infatti, essere utilizzato per confrontare modi diversi di attuarne aspetti minori. Come molti degli altri esempi illustrati in questo documento, una delle cose più rilevanti degli RCT è la possibilità di testare l'efficacia di particolari aspetti di un programma più ampio. Testare piccole parti di un programma permette al decisore di raffinare continuamente la politica, focalizzandosi sugli elementi che producono il maggiore impatto.

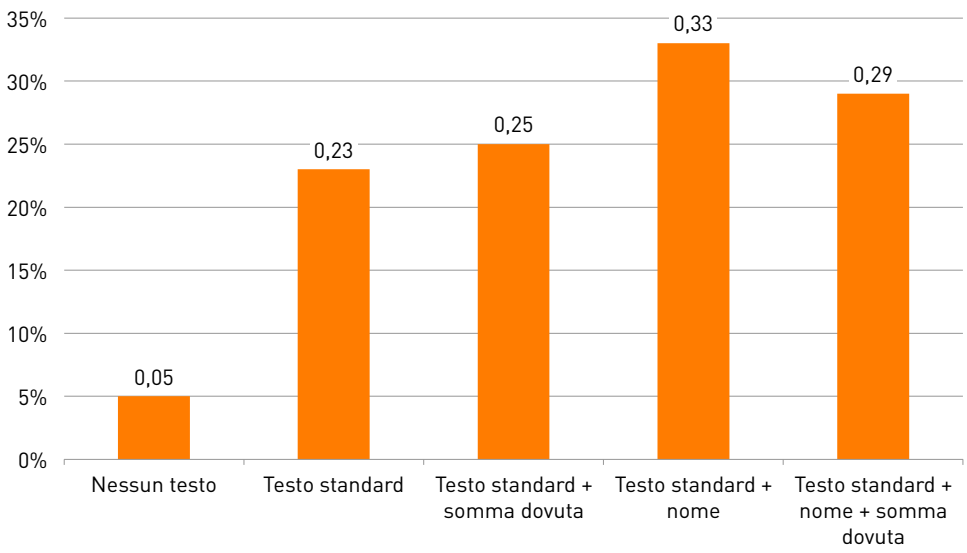
Sia per il confronto di due interventi completi, sia di specifici aspetti di una politica, valgono gli stessi principi di base dell'esperimento randomizzato: confrontando due gruppi di beneficiari identici, scelti a caso, è possibile controllare gli effetti di molte variabili di intervento, così da capire cosa funziona e cosa non funziona (riquadri 2, 3 e 4).

Riquadro 2 - L'impatto dell'invio di SMS per sollecitare il pagamento delle contravvenzioni

Il *Courts Service** e il *Behavioural Insights Team* hanno voluto verificare se l'invio di SMS a coloro che non erano riusciti a pagare le contravvenzioni li avrebbe incoraggiati a pagare prima di ricevere la visita di un ufficiale giudiziario. Per rispondere a questa domanda si è utilizzato l'approccio "sperimenta, impara, adatta" che prevede la sperimentazione simultanea di una serie di varianti per scoprire quella che funziona meglio.

Nello studio iniziale, gli individui sono stati assegnati in modo casuale a cinque diversi gruppi. A un gruppo (di controllo) non è stato inviato nessun messaggio, mentre agli altri gruppi (di intervento) sono stati inviati un testo standard promemoria o messaggi più personali (che includevano il nome del destinatario, l'importo dovuto o entrambe le cose). L'esperimento ha dimostrato che l'invio di SMS può essere molto efficace (figura 1.3).

Figura 1.3 - Prova iniziale: tassi di rimborso da parte degli individui (N = 1.054)



I dati rappresentano i tassi di risposta per tipo di messaggio inviato (ai soggetti con n. di cellulare corretto)

In seguito - su un campione più ampio di individui (N = 3.633) - è stato condotto un secondo esperimento teso a individuare gli aspetti dei messaggi personalizzati che più hanno contribuito ad accrescere i tassi di pagamento. Il quadro dei risultati ottenuti è assai simile al primo esperimento. Tuttavia, in questo caso l'esperimento non solo ha permesso di confermare che le persone sono più propense pagare la propria contravvenzione quando ricevono un SMS contenente il loro nome ma anche che, in questo modo, il valore medio dei pagamenti cresce di oltre il 30%.

I due studi sono stati condotti sostenendo costi molto bassi: i dati dei pagamenti erano già raccolti dal *Courts Service*, mentre l'unico costo aggiuntivo era costituito dal tempo dei ricercatori per impostare l'esperimento. Se implementata a livello nazionale, la personalizzazione del messaggio aumenterebbe radicalmente gli incassi delle multe non pagate. Si stima infatti che, inviando un testo personalizzato piuttosto che un testo *standard*, si raccoglierebbero oltre £3 milioni all'anno in più. I maggiori incassi conseguibili dall'utilizzo di testi personalizzati supererebbero

dunque di molte volte il costo dell'invio. In aggiunta a ciò, il *Courts Service* stima che l'invio di solleciti personalizzati potrebbe ridurre la necessità di almeno 150.000 interventi degli ufficiali giudiziari ogni anno.

* L'Her Majesty Courts Service (HMCS) era un'Agenzia esecutiva del Ministero della Giustizia (MOJ) responsabile per l'amministrazione dei tribunali civili, familiari e penali di Inghilterra e del Galles.

Riquadro 3 - L'uso degli studi randomizzati controllati per valutare cosa funziona nelle politiche di reinserimento al lavoro

Nel 2003, il Ministero britannico del lavoro e delle pensioni (DWP) ha condotto uno studio randomizzato controllato per valutare l'impatto di tre nuovi programmi per i richiedenti delle misure di *Incapacity Benefit* (DWP, 2006b): sostegno al lavoro; supporto sanitario o entrambe*. Il costo del sostegno supplementare era in media di £1.400 ma lo studio non ha rilevato benefici aggiuntivi rispetto al supporto tradizionale, già disponibile. Lo studio ha fatto quindi risparmiare al contribuente molti milioni di sterline fornendo la prova incontrovertibile che il nuovo programma non produceva gli effetti attesi.

Più di recente, lo stesso Ministero ha tenuto a verificare se si potessero rendere meno severe le procedure di accesso al beneficio senza peggiorare i risultati. Lo studio ha coinvolto oltre 60.000 persone in cerca di lavoro, il tradizionale processo di richiesta quindicinale del beneficio è stato confrontato con molti altri metodi che richiedevano uno sforzo minore (ad esempio, la conferma telefonica o meno frequente). Tutte le alternative allo *status quo* testate - in studi di dimensioni sufficienti a produrre risultati statisticamente significativi - hanno mostrato un allungamento del tempo necessario ai beneficiari per trovare un lavoro (DWP, 2006a). Di conseguenza, nonostante le varie modifiche introdotte all'*Incapacity Benefit*, il Ministero ha continuato a richiedere la conferma dell'iscrizione su base quindicinale.

* Si tratta di un disegno d'interazione che consente la determinazione degli effetti separati e combinati di due interventi. Tale metodologia è particolarmente utile qualora s'intendano valutare gli effetti addizionali di una o più caratteristiche di un programma complesso.

L'*Incapacity Benefit* era un sussidio sociale introdotto nel Regno Unito nel 1995 dal governo Major. Ne avevano diritto le persone di età inferiore all'età pensionabile in regola con i contributi e in difficoltà a trovare lavoro a causa di una malattia o una disabilità. Il numero dei destinatari è aumentato notevolmente dal 1995 al 2004. Ciò ha messo in discussione la severità del processo utilizzato per testare l'idoneità, i rischi di generare "trappole" per i beneficiari e preoccupazioni sulla sostenibilità dei costi per l'erario. La misura è stata gradualmente eliminata dopo l'approvazione del *Welfare Reform Act* del 2007. Nel 2011, il governo conservatore-liberaldemocratico ha cominciato a rivalutare i requisiti dei beneficiari di lungo periodo per escludere quelli effettivamente in grado di lavorare (Nota del traduttore).

Riquadro 4 - Il legame tra teorie della crescita, innovazione ed esperimenti randomizzati controllati

Il crescente interesse per l'uso di esperimenti randomizzati come strumento di *policy making* è coerente con le più recenti correnti di pensiero. Quando le risorse sono scarse è essenziale assicurarsi che vengano spese per implementare politiche che funzionano e anche piccoli miglioramenti marginali nel rapporto costo-efficacia sono preziosi. Gli esperimenti controllati sono uno strumento estremamente potente per individuare il costo-efficacia e tagliare le spese meno produttive.

L'espressione pratica di questo pensiero spinge verso un maggiore decentramento del processo decisionale e un più ampio sfruttamento del mercato per fornire beni e servizi. La generazione di innovazione deve essere accompagnata da meccanismi che identifichino e alimentino le innovazioni di successo. Ciò include una maggiore trasparenza e lo sviluppo di meccanismi di *feedback* sia nei mercati dei prodotti di consumo sia in quelli dei servizi pubblici, in grado di generare prestazioni migliori e, spesso, la crescita di produttori più piccoli e indipendenti (Luca, 2011). Nei servizi pubblici, e ovunque i meccanismi di mercato e di pagamento a fronte di risultati possano essere inappropriati, le sperimentazioni controllate e su più rami possono svolgere un ruolo molto importante, soprattutto se i risultati sono ampiamente diffusi e applicati.

2

2. SFATIAMO ALCUNI MITI SUGLI STUDI RANDOMIZZATI CONTROLLATI

Ci sono molti campi nei quali gli studi randomizzati sono ormai prassi comune e non farli sarebbe considerato bizzarro o, addirittura, imprudente. Tali studi sono infatti il metodo universalmente riconosciuto in medicina: quando si tratta di capire quale tra due trattamenti medici funziona meglio, l'efficacia di un nuovo farmaco o di differenti tipologie di chirurgia anti cancro o, addirittura, di diversi modelli di calze contenitive. La realtà non è però sempre stata questa: quando le prove furono introdotte per la prima volta in medicina trovarono una forte opposizione in alcuni clinici, molti dei quali credevano che il loro giudizio personale "esperto" fosse sufficiente per decidere dell'efficacia di un particolare trattamento.

Gli studi randomizzati sono però sempre più utilizzati anche per studiare l'efficacia e il rapporto costo-efficiacia di vari programmi internazionali di sviluppo (riquadro 5). Anche nel mondo degli affari - ad esempio quando le aziende vogliono scoprire quale tra due differenti pagine *web* è in grado di produrre più *click-through*[3] e vendite - è comune assegnare in modo casuale i visitatori a differenti versioni dei siti *web* monitorandone i percorsi e i comportamenti di acquisto (riquadro 6).

Anche se si trovano alcuni buoni esempi di utilizzo degli studi randomizzati da parte dei *policy maker*, questi non sono ancora così diffusi in questo campo. Ciò può essere in parte dovuto a una mancanza di consapevolezza ma anche a molti malintesi che portano spesso a rifiuti impropri.

In questa sede passiamo in rassegna ciascuno di questi miti in modo da rispondere all'accusa, infondata, che gli RCT sono sempre difficili, costosi, contrari all'etica o non necessari. Gli autori ritengono, al contrario, che il pericolo sia insito piuttosto nell'eccessiva sicurezza nel presumere che gli interventi siano efficaci e che gli studi randomizzati svolgano un ruolo fondamentale per dimostrare non solo l'effettiva efficacia di un intervento ma anche il suo rapporto costo-efficacia.

3 Il *click-through rate* (percentuale di click) è un tasso che misura l'efficacia di una campagna pubblicitaria *on-line*.

Riquadro 5 - Studi randomizzati controllati per migliorare i risultati educativi in India

Una delle aree di più rapida crescita nell'uso degli esperimenti randomizzati è stata, negli ultimi anni, la cooperazione internazionale. In questo campo sono stati realizzati numerosi studi per determinare i modi più appropriati per sconfiggere la povertà nei paesi in via di sviluppo: da quelli per aumentare le rese delle colture, a quelli per incoraggiare l'uso delle zanzariere, assicurare la presenza degli insegnanti in classe, promuovere l'imprenditorialità o aumentare i tassi di vaccinazione.

Ad esempio, lo sforzo effettuato negli ultimi decenni per rendere l'istruzione universalmente disponibile nei paesi in via di sviluppo ha portato a un aumento delle iscrizioni e della frequenza scolastica. Tuttavia, la qualità dell'istruzione a disposizione dei bambini provenienti dalle famiglie più povere rimane un problema grave: un'indagine effettuata del 2005 in India ha mostrato che oltre il 40% dei bambini sotto i 12 anni non era in grado di leggere un testo semplice e il 50% non riusciva a eseguire una semplice sottrazione. In collaborazione con un'ONG attiva nel campo dell'educazione, alcuni ricercatori americani hanno quindi condotto uno studio randomizzato per valutare gli effetti di una riduzione dei costi dei corsi di recupero sui risultati scolastici. L'esperimento ha coinvolto circa 200 scuole cui sono stati assegnati, in modo casuale, dei tutor per assistere gli studenti iscritti alla terza o alla quarta classe. L'impatto del programma è stato misurato confrontando i risultati scolastici degli studenti delle terze classi con e senza il tutor.

I tutor erano donne della comunità locale pagate con una somma modesta (una frazione dello stipendio di un maestro) e hanno lavorato separatamente, per metà della giornata scolastica, con gruppi di bambini rimasti indietro nel programma. I risultati hanno dimostrato che il progetto di recupero migliora in modo significativo i punteggi nei test, in particolare di matematica (Banerjee *et al.*, 2007). L'intervento si è rivelato un tale successo (e con un rapporto costo-efficacia migliore rispetto ad altri programmi rivolti ad aumentare la *performance* scolastica) che è stato esteso a tutta l'India.

2.1 Non sempre sappiamo cosa funziona

Policy maker e operatori sono spesso convinti di sapere quali interventi funzionano meglio e utilizzano queste convinzioni per elaborare le politiche. Anche se ci sono fondati motivi per ritenere che una politica sarà efficace, vale sempre la pena con-

Riquadro 6 - L'utilizzo degli studi randomizzati per migliorare le *performance* d'impresa

Molte aziende utilizzano sempre più diffusamente esperimenti randomizzati per testare la risposta dei consumatori a diverse presentazioni di prodotti *on-line*. Poche di queste informazioni sono a disposizione del pubblico ma è ben noto che aziende come Amazon ed eBay utilizzano il traffico *web* di *routine* sui loro siti per testare cosa funziona meglio per stimolare gli acquisti. Ad esempio, i clienti potrebbero essere indirizzati a versioni differenti di una pagina web. Tracciando i *click-through* e i comportamenti di acquisto dei clienti che transitano nelle diverse versioni del sito web, le aziende possono modificarne disegno e configurazione in modo da massimizzare i profitti. Seguono alcuni ulteriori esempi.

Durante la recente raccolta di fondi, una foto del fondatore di *Wikipedia*, Jimmy Wales, è apparsa nella pubblicità delle donazioni nella parte superiore della pagina: il metodo è stato scelto utilizzando i risultati di una serie di studi che confrontavano la propensione a donare dei visitatori del sito a seconda del modello di pubblicità che veniva loro proposto in modo casuale.

Netflix, società che offre film in *streaming on-line*, effettua spesso esperimenti simultanei sull'esperienza dei propri utenti. Per sperimentare la *Netflix Screening Room*, un nuovo modo di visione dei film in anteprima, sono state sviluppate quattro diverse versioni del servizio. Queste sono state offerte a quattro gruppi di 20.000 abbonati e a un gruppo di controllo che ha continuato a ricevere il normale servizio di Netflix. Il comportamento degli utenti è stato monitorato per confrontare la propensione alla visione nei differenti gruppi (Davenport, Harris, 2007).

Delta Airlines ha utilizzato la sperimentazione randomizzata per migliorare il proprio sito *web*. Nel 2006, nonostante la generale crescita di prenotazioni di viaggi *on-line*, il traffico *web* al sito Delta Airlines non riusciva a generare il numero atteso di prenotazioni. Quasi il 50% dei visitatori del sito rinunciava prima di completare il processo di prenotazione: dopo aver selezionato il proprio volo, i potenziali clienti abbandonavano spesso il sito quando raggiungevano la pagina che richiedeva l'inserimento dei dati personali (nome, indirizzo, dati della carta). Invece di cambiare l'intero sito, Delta si è concentrata sulle pagine specifiche che non riuscivano a convertire i potenziali clienti in vendite. A questo scopo, numerose varianti di tali pagine sono state testate *on-line* assegnando casualmente i clienti. Delta ha quindi scoperto che, eliminando le istruzioni dettagliate pubblicate nella parte superiore della pagina di richiesta dei dati personali, i clienti erano molto più propensi a portare a termine il processo di prenotazione. Applicando questo e altri cambiamenti minori identificati durante il *test*, i tassi di conversione per le vendite di biglietti sono migliorate del 5% (Delta Airlines, 2007), un cambiamento piccolo ma di grande valore economico.

durre un esperimento randomizzato per quantificare accuratamente il beneficio che si è in grado di produrre. Uno studio può anche contribuire a individuare quali aspetti di un programma stanno producendo il massimo effetto e come questo potrebbe essere ulteriormente migliorato. Per esempio, se stiamo implementando un nuovo programma di finanziamento di *start-up* innovative, sarebbe utile sapere se raddoppiare i fondi a disposizione ha un effetto significativo sul successo dell'iniziativa o non fa alcuna differenza.

Dobbiamo anche riconoscere che spesso le previsioni degli esperti si rivelano sbagliate. Gli studi randomizzati hanno dimostrato che interventi considerati efficaci in realtà non lo erano (riquadro 3). Altre volte si è invece dimostrato il contrario: interventi sui quali vi era uno scetticismo iniziale sono, in ultima analisi, risultati efficaci. Ad esempio, quando il *Behavioural Insights Team* e il *Courts Service* hanno valutato l'effetto dei messaggi sulla propensione a pagare le multe, pochi avevano inizialmente previsto che il testo personalizzato avrebbe aumentato in modo significativo i tassi di rimborso (riquadro 2).

Ci sono però anche innumerevoli esempi di esperimenti che hanno ribaltato ipotesi consolidate su ciò che funziona, mostrando che interventi ritenuti efficaci erano, in realtà, dannosi. Il caso dell'iniezione di steroidi (riquadro 7) è un potente esempio di come le ipotesi non necessariamente superano la prova empirica. Allo stesso modo, il programma *Scared Straight*, che espone i giovani alla realtà di una vita criminale, è un buon esempio di una politica basata sulle migliori intenzioni e fondata su una base apparentemente consistente di evidenze ma che, alla prova di un esperimento randomizzato, ha mostrato di produrre effetti negativi (riquadro 8). Gli studi randomizzati sono il migliore metodo disponibile per evitare questi errori, perché forniscono ai *policy maker* e agli operatori una prova solida dell'efficacia di un intervento comparandolo con quanto sarebbe successo in sua assenza (riquadro 9).

Riquadro 7 - L'utilizzo degli steroidi nei casi di trauma cranico: salvare o uccidere le persone?

Per diversi decenni, gli adulti con trauma cranico grave sono stati trattati con iniezioni di steroidi. Ciò aveva perfettamente senso in linea di principio: gli steroidi riducono il gonfiore e quindi si credeva che potesse proteggere le persone con ferite alla testa dallo schiacciamento del cervello. Tuttavia, queste ipotesi per moltissimo tempo non sono state sottoposte a *test* adeguati.

Una decina di anni fa, questa ipotesi è stata finalmente testata con uno studio randomizzato. Lo studio è stato controverso, e molti vi si opposero perché erano certi dell'efficacia degli steroidi. Quando i risultati dello studio furono pubblicati nel 2005 (Edwards *et al.*, 2005), dimostravano che le persone curate con iniezioni di steroidi avevano probabilità maggiori di morire: di fatto, il trattamento di routine stava uccidendo un gran numero di persone perché le lesioni alla testa sono molto comuni. Dati i risultati, anche il *test* fu immediatamente interrotto per evitare ulteriori danni.

Questo è un esempio particolarmente drammatico del perché siano importanti prove eque di interventi nuovi ed esistenti: senza di loro, potremmo involontariamente infliggere un danno, senza mai venirne a conoscenza; e quando nuovi interventi diventano pratica comune senza aver superato *test* rigorosi, si potrebbe anche manifestare una resistenza al loro successivo collaudo.

2.2 Gli studi randomizzati controllati non sono necessariamente costosi

I costi di uno studio randomizzato dipendono da come questo è stato progettato: con una corretta pianificazione i costi possono rivelarsi inferiori rispetto ad altre forme di valutazione. Ciò è particolarmente vero quando un servizio è già in opera e i dati sui risultati vengono già raccolti da sistemi di monitoraggio di *routine*, come avviene in molte parti del settore pubblico (dati amministrativi). Contrariamente a quanto avviene con gli studi clinici, un esperimento di politica pubblica non necessariamente richiede di reclutare i partecipanti al di fuori della pratica consueta o di mettere in atto nuovi sistemi per fornire interventi o monitorarne i risultati.

Il *Behavioural Insights Team* ha lavorato con molti ministeri per eseguire esperimenti i cui costi aggiuntivi erano solo quelli strettamente necessari a coprire il costo del lavoro dei membri del *team*. Ad esempio, negli studi eseguiti con le autorità locali, l'HMRC[4], il DVLA[5], il *Courts Service*[6] e il Job Centre Plus[7] la sperimentazione - dell'invio di una lettera, di un servizio di consulenza per i disoccupati, etc. - ha utilizzato le strutture e i processi di erogazione di servizi e rilevazione dei dati già in uso.

Nel calcolare le risorse aggiuntive necessarie all'esecuzione di un esperimento occorre ricordare che si tratta spesso del modo migliore per valutare il costo-efficacia di un intervento. In alcuni casi, un esperimento consente di capire che un programma è troppo costoso rispetto ai benefici che genera. In altri casi, l'esperimento può dimostrare un eccellente rapporto costi-efficacia di un programma che meriterebbe di essere ulteriormente esteso.

Mostrando quanto l'intervento sperimentato si è rivelato più o meno efficace di quello tradizionale, i suoi responsabili sono anche in grado di giudicare se il suo costo è effettivamente giustificato dai benefici generati. Invece di considerare il costo della sperimentazione, potrebbe quindi essere più opportuno valutare i costi della sua mancata realizzazione[8].

2.3 Ci sono vantaggi etici per ricorrere agli esperimenti

A volte le persone rifiutano le sperimentazioni perché sostengono che sarebbe immorale privare del nuovo intervento persone che ne potrebbero trarre un beneficio. Questo succede in particolare quando l'intervento potrebbe migliorare la salute, la ricchezza o il livello di istruzione di un gruppo.

È sicuramente vero che può essere difficile escludere qualcuno dai benefici di un intervento, ma è bene ricordare che, molto spesso, quello che si sta sperimentando

4 L'*Her Majesty's Revenue and Customs* (HMRC) è l'Agenzia delle entrate e delle dogane del Regno Unito.

5 La *Driver and Vehicle Licensing Agency* (DVLC) del Regno unito è l'agenzia omologa dell'Ispettorato generale della motorizzazione civile e dei trasporti in concessione.

6 Il *Courts Service* era l'Agenzia esecutiva del Ministero della Giustizia (MOJ) responsabile per l'amministrazione dei tribunali civili, familiari e penali.

7 Il *Jobcentre Plus* era un'agenzia esecutiva del Ministero del lavoro (DWP) del Regno Unito attiva tra il 2002 e il 2011.

8 Questo costo può essere stimato comparando il costo di una sperimentazione, ad esempio, con la quantità di denaro che sarebbe sprecata se l'intervento implementato non mostrasse alcun beneficio.

non ha ancora dimostrato di essere utile. Chi scrive ritiene che si debba essere molto chiari sui limiti della conoscenza e che quindi non si possa essere certi dell'efficacia di un intervento fino a quando questo non sarà valutato con metodi scientificamente robusti.

A volte, anche interventi ritenuti efficaci si sono rivelati inefficaci o addirittura dannosi (riquadri 7 e 8). Questo può accadere anche con politiche delle quali sarebbe ragionevole presumere una garanzia di efficacia. Ad esempio, per incoraggiare soggetti adulti a partecipare a corsi di alfabetizzazione sono stati spesso utilizzati degli incentivi ma, valutando gli effetti con uno studio randomizzato, si è riscontrato che i partecipanti che ricevevano gli incentivi partecipavano alle lezioni in misura minore rispetto ai soggetti non incentivati (Brooks *et al.*, 2008). Questo studio ha quindi dimostrato non solo che gli incentivi costituivano uno spreco di risorse ma che riducevano anche la frequenza alle lezioni. Cancellare l'intervento era quindi meglio che continuarlo e, se non si fosse condotto l'esperimento, gli studenti avrebbero continuato a essere danneggiati, nonostante le migliori intenzioni e senza saperlo.

È anche interessante notare che le politiche sono spesso implementate in maniera graduale o a scaglioni, con alcune regioni anticipatrici; questa gradualità non è generalmente considerata immorale. La fornitura del programma *Sure Start* è un esempio di questo tipo. Un'introduzione graduale dell'intervento utilizzata nell'ambito di un esperimento è più etica, perché genera nuove e importanti informazioni che permettono di comprendere se un intervento è economicamente vantaggioso.

2.4 Gli studi randomizzati non sono sempre complicati da realizzare

Gli studi randomizzati nella loro forma più semplice sono molto semplici da eseguire. Tuttavia ci possono essere alcune insidie nascoste che rendono preferibile, almeno all'inizio, fare ricorso a un po' di competenze specifiche. Alcune di queste "trappole" saranno descritte nel prossimo capitolo ma, in linea di massima, non sono più impegnative di quelle che si incontrano in qualsiasi altra attività di valutazione dei risultati e possono essere superate con gli strumenti adeguati.

Lo sforzo iniziale per costruire un esperimento randomizzato e definire chiaramente i risultati prima dell'avvio del progetto è sempre tempo ben speso. In mancanza di uno studio randomizzato, infatti, ogni tentativo di provare e valutare l'impatto di un intervento sarà difficile, costoso e produrrà risultati meno oggettivi e più distorti, dato che sarà più difficile attribuire gli effetti osservati all'intervento e non alle altre possibili cause esterne. In generale, vale la pena dedicare qualche risorsa in più e progettare un disegno di valutazione randomizzato prima dell'avvio di qualsiasi politica.

Riquadro 8 - Il programma *Scared Straight!*: prevenzione o incoraggiamento della delinquenza minorile?

Scared Straight! è un programma sviluppato negli Stati Uniti per scoraggiare i comportamenti criminali di giovani delinquenti e dei bambini a rischio. Il programma prevede l'esposizione dei minori alle spaventose conseguenze che derivano da una condotta criminale facendoli interagire con pericolosi criminali in custodia.

La teoria sottostante al programma è che i bambini sarebbero meno propensi a delinquere se messi a conoscenza delle gravi conseguenze che ne possono derivare. Diversi studi preliminari che hanno osservato i comportamenti criminali dei partecipanti prima e dopo il programma, sembravano confermare la validità della teoria (Finckenauer, 1982). I tassi di successo registrati erano così elevati (fino al 94%) che il programma è stato adottato in diversi paesi, tra i quali il Regno Unito.

Nessuna delle valutazioni effettuate aveva però utilizzato un gruppo di controllo in grado di mostrare cosa sarebbe successo ai partecipanti se non avessero preso parte al programma. Diversi studi randomizzati hanno quindi cercato di porre rimedio al problema. Una meta-analisi di 7 studi statunitensi che hanno assegnato al programma in modo casuale la metà del campione dei bambini a rischio selezionati ha mostrato che il programma *Scared Straight!* generava tassi più elevati di comportamenti criminali: "meglio non fare nulla che esporre i minori al programma" (Petrosino *et al.*, 2003). Analisi più recenti suggeriscono altresì che i costi associati al programma (in gran parte legati all'aumento dei tassi di recidiva) risultano oltre 30 volte superiori ai benefici, il che significa che il programma genera sia un aumento del crimine (The Social Research Unit, 2012), sia un aggravio dei costi per il contribuente.

Riquadro 9 - Il *Family Nurse Partnership* (Partenariato famiglia - infermiera): una valutazione rigorosa per la diffusione del progetto

Il *Family Nurse Partnership* (FNP) è un programma di prevenzione per madri primipare vulnerabili. Sviluppato negli Stati Uniti, prevede una serie intensiva di visite a domicilio svolte da infermieri appositamente addestrati per offrire supporto a partire dalla prime fasi della gravidanza fino ai due anni di vita del bambino. Molti studi randomizzati effettuati negli Stati Uniti* hanno rilevato benefici significativi per le giovani famiglie svantaggiate e notevoli risparmi di costi. Ad esempio, i bambini delle famiglie beneficiarie mostrano un migliore sviluppo socio-emotivo, migliori risultati scolastici e minore probabilità di essere coinvolti in attività criminali. Le madri sono soggette a un minor numero di gravidanze successive e intervalli più lunghi tra le nascite, hanno maggiori possibilità di impiego e minori di delinquere.

Il programma è stato adottato nel Regno Unito a partire dal 2007, spesso attraverso gli *Sure Start Centres* per bambini e nel 2015 il Ministero della salute si è impegnato a raddoppiare il numero di giovani madri destinatarie dell'intervento portandolo a 13.000 (l'intervento viene erogato una sola volta, alla nascita del primo figlio). Nel frattempo, il Ministero sta finanziando una valutazione randomizzata del programma, per valutare se benefici e costo-efficacia superano quelli dei servizi tradizionali. La sperimentazione si svolge in 18 siti sparsi in tutto il Regno Unito, coinvolgendo circa 1.650 donne: si tratta del più ampio esperimento mai avviato su questo programma. Tra i risultati rilevati si trovano il fumo durante la gravidanza, l'allattamento al seno, i ricoveri in ospedale per lesioni e ingestione, ulteriori gravidanze e lo sviluppo del bambino a 2 anni.

* Per una rassegna delle ricerche statunitensi sul programma si veda: MacMillan, *et al*. (2009).

3

3. COME SI REALIZZA UNO STUDIO RANDOMIZZATO CONTROLLATO?

La prima parte di questo documento ha discusso dell'utilizzo degli studi randomizzati nelle politiche pubbliche. La seconda parte riguarda invece le sue modalità operative. Senza cercare di essere esaustivi, in questi capitoli si delineano i passi necessari per ogni esperimento e si indicano le aree nelle quali un *policy maker* potrebbe aver bisogno del supporto di esperti metodologi. Indichiamo in primo luogo le nove fasi fondamentali di ogni studio randomizzato. Molti di queste fasi sono comuni a qualsiasi disegno di valutazione di una politica ben progettata - per esempio, la necessità di essere chiari, fin dall'inizio, su ciò che la politica sta cercando di perseguire. Altri elementi saranno invece meno noti, in particolare la necessità di prevedere diversi gruppi di popolazione composti in modo casuale e ai quali saranno assegnate versioni differenti dell'intervento. Queste fasi sono riassunte nel riquadro 10 e illustrate dettagliatamente in seguito.

Riquadro 10 – Le fasi di lavoro

Sperimenta

1. Identifica due o più interventi da comparare (ad esempio una vecchia e una nuova politica; diverse varianti di una politica).
2. Determina esattamente il risultato (*outcome*) che la politica si propone di ottenere e come sarà misurato durante l'esperimento.
3. Decidi l'unità di randomizzazione nei gruppi d'intervento e di controllo: individui, istituzioni (ad esempio scuole) o aree geografiche/amministrative (ad esempio enti locali).
4. Determina il numero di unità (persone, istituzioni o aree) necessarie per ottenere risultati statisticamente significativi.
5. Assegna le unità di randomizzazione agli interventi utilizzando un metodo di randomizzazione robusto.
6. Avvia gli interventi per i destinatari assegnati.

Impara

7. Misura i risultati e determina l'effetto degli interventi.

Adatta

8. Adatta il tuo intervento tenendo conto dei risultati ottenuti.
9. Ritorna al punto 1 per continuare a migliorare.

3.1 Sperimenta

Fase 1 – Identificare due o più interventi da confrontare

Gli studi randomizzati si svolgono quando c'è incertezza nell'identificazione del migliore tra due o più interventi alternativi mettendoli direttamente a confronto tra loro. Spesso, gli studi servono a confrontare un nuovo intervento con la pratica tradizionale. Il nuovo intervento potrebbe consistere in una sola piccola variante o in un insieme di piccole modifiche alla prassi attuale (riquadri 11 e 12), oppure potrebbe essere un approccio completamente nuovo che si sta dimostrando efficace in un altro paese o in un contesto diverso o che ha basi teoriche molto solide.

Prima di progettare uno studio randomizzato è importante considerare quanto è già noto circa l'efficacia dell'intervento che ci si propone di testare. In alcuni casi potrebbero, per esempio, essere disponibili risultati di studi randomizzati condotti in contesti simili. Ricorrere a ricerche esistenti può anche aiutare a sviluppare o migliorare l'intervento. Un buon punto di partenza da cui partire è quello della Campbell Collaboration[9] che aiuta *policy makers* e professionisti raccogliendo e sistematizzando l'evidenza scientifica prodotta sulle politiche pubbliche.

È anche importante che le sperimentazioni siano condotte su interventi effettivamente implementabili laddove l'esperimento ne mostrasse l'efficacia. Spesso, infatti, c'è la tentazione di eseguire un esperimento randomizzato su una politica ideale, i cui costi non sarebbero però sostenibili una volta a regime. In un caso di questo tipo lo studio randomizzato sarebbe del tutto inutile perché non si sarebbe in grado di generalizzare i risultati mancando le risorse per estendere la politica su scala nazionale.

Dobbiamo anche essere sicuri che i risultati dello studio saranno effettivamente replicabili nel caso in cui la politica dovesse essere diffusa più ampiamente. Perché i risultati siano generalizzabili e rilevanti, l'intervento, gli operatori che lo implementano e i dati che ne descrivono i risultati, dovranno essere quanto più possibile rappresentativi della realtà. Quando il *Behavioural Insights Team* conduce studi randomizzati per la valutazione delle politiche pubbliche, il gruppo di lavoro rimane per un certo periodo a stretto contatto con le organizzazioni sul campo, in modo da capire cosa sia effettivamente fattibile e se il personale abbia eventualmente già sviluppato interventi alternativi - potenzialmente efficaci ma non ancora testati - per il raggiungimento degli stessi risultati.

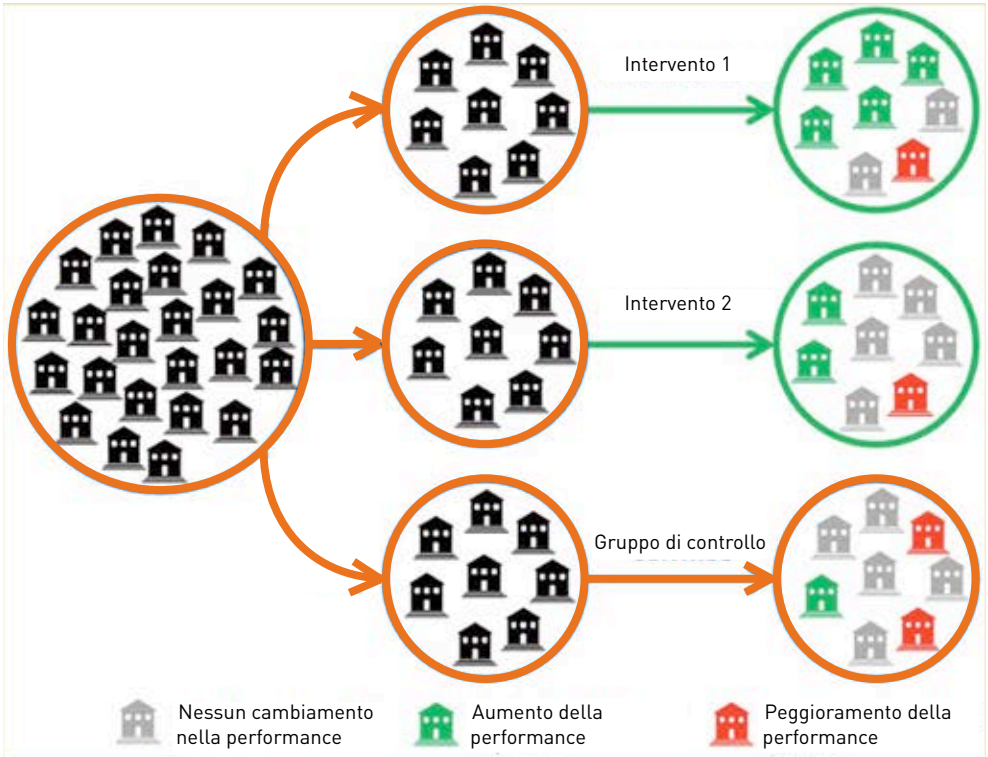
9 www.campbellcollaboration.org.

Riquadro 11 - Confrontare diverse opzioni di una politica e testare piccole variazioni

Un esperimento randomizzato non confronta necessariamente l'intervento contro il nulla. Molti interventi potrebbero rivelarsi probabilmente più utili rispetto al non fare niente. Gli esperimenti possono invece essere utilizzati più proficuamente per identificare l'opzione migliore tra alcune di quelle disponibili.

In alcuni casi, potremmo essere interessati a identificare l'opzione di *policy* più appropriata. Immaginiamo, per esempio, di avere risorse che potrebbero essere dedicate a una sola di due opzioni differenti: migliorare le strutture informatiche in tutte le scuole secondarie oppure pagare più insegnanti. In un caso come questo, potremmo eseguire un esperimento articolato in 3 diversi rami (figura 6): un gruppo di controllo (costituito da alcune scuole che continuano con le infrastrutture informatiche attuali e lo stesso numero di insegnanti) e due gruppi di intervento (scuole che ricevono l'aggiornamento informatico e scuole che ricevono più insegnanti). Questo ci permetterebbe di stabilire quale intervento si sia rivelato più efficace e abbia offerto il migliore rapporto costo-efficacia.

Figura 3.1 - Il disegno di un ipotetico studio randomizzato multi-ramo per valutare il miglioramento del rendimento scolastico degli studenti generato da due tipi di intervento differenti: l'aggiornamento dell'infrastruttura informatica delle scuole (intervento 1) o l'aumento del numero degli insegnanti (intervento 2)



In altri casi, potremmo invece essere interessati a rispondere a domande più sottili su una politica particolare, come, ad esempio, quali piccole variazioni di dettaglio generano i risultati migliori. Ad esempio, immaginiamo di dover introdurre alcune modifiche al menù servito nelle mense scolastiche. Potrebbe essere già corso un tentativo di introdurre alcune opzioni alimentari più sane allo scopo di migliorare le diete dei bambini. Pur sapendo l'importanza della presentazione del cibo, potremmo non conoscere quale sia il modo più efficace per presentare il cibo ai bambini e incoraggiarli a mangiare più sano. Potremmo quindi eseguire uno studio multi ramo col quale offrire diverse varianti di presentazione (l'ordine di insalate e cibi caldi, la dimensione delle siviere e dei piatti, etc.).

Spesso, l'opportunità per l'affinamento delle politiche sorge proprio quando stanno per essere introdotti dei cambiamenti: si tratta infatti del momento ideale per testare qualche lieve variazione e garantire che i cambiamenti introdotti ottengano l'effetto desiderato.

Fase 2 - Definire con precisione il risultato e la sua misurazione

All'inizio di ogni esperimento è fondamentale definire esattamente il risultato atteso e il modo per misurarlo. Ad esempio, nell'ambito di una politica educativa, una misura di esito potrebbe essere costituita dai risultati degli studenti in sede di esame. Nell'ambito delle politiche per il miglioramento dell'efficienza energetica, il risultato potrebbe essere misurato dal consumo di energia delle famiglie.

Sempre in fase di progettazione, è importante precisare anche le modalità e i tempi di misurazione dei risultati, mantenendoli anche in fase di analisi. Naturalmente è fondamentale garantire che la misurazione dei risultati sia esattamente la stessa per tutti i gruppi, sia in termini di modalità di misurazione, sia degli standard applicati.

Definire in anticipo le misure di risultato non solo è una cosa di buon senso: ci sono, infatti, anche buone ragioni scientifiche che rendono tale operazione cruciale per il successo di un esperimento. Nel corso del tempo, si verificheranno infatti sempre delle oscillazioni casuali dei dati raccolti. Al termine dell'esperimento, saranno disponibili molti dati su molte cose diverse e sarà inevitabile che alcuni di essi mostreranno miglioramenti - o peggioramenti - per motivi del tutto casuali.

Ogni volta che si verifica una simile variazione casuale, si potrebbe essere tentati di scegliere numeri che mostrano per puro caso un miglioramento e utilizzarli come prova del successo dell'iniziativa. In questo modo però s'infrangerebbero le ipotesi dei *test* statistici utilizzati nell'analisi dei dati, perché ci daremmo troppe possibilità di trovare un risultato positivo. La tentazione di sovrastimare i dati, attribuendo all'intervento anche variazioni casuali, si può evitare solo specificando in anticipo i risultati da osservare. I *test* statistici servono proprio per capire quanta parte della variazione riscontrata sia attribuibile semplicemente al caso.

Nel decidere il risultato da misurare, è poi importante scegliere una grandezza realmente importante o una sua fedele approssimazione, piuttosto che un dato riferito al processo, che fornirebbe indicazioni assai differenti. Ad esempio, in uno studio per la

Riquadro 12 - Cogliere le opportunità disponibili per realizzare studi randomizzati

A volte, alcuni vincoli alla realizzazione delle politiche offrono il contesto ideale per l'esecuzione di loro sperimentazioni. Ad esempio, vincoli finanziari e/o operativi possono determinare uno sfalsamento su scala territoriale dell'implementazione della politica. Nella misura in cui fosse possibile monitorare i risultati in tutte le aree che saranno sottoposte all'intervento e decidere in modo casuale l'ordine con il quale queste lo riceveranno, tale sfalsamento potrà essere sfruttato per eseguire un disegno valutativo a stadi (*stepped-wedge*).

Ad esempio, il servizio di messa alla prova (libertà vigilata) nella zona di Durham (GB) ha voluto sperimentare un nuovo approccio. I vincoli nella dotazione di risorse hanno impedito la possibilità di realizzare il programma di orientamento e formazione contemporaneamente in tutti i 6 centri presenti sul territorio. Il modo più equo e scientificamente robusto per procedere è stata quindi la costruzione casuale dell'ordine con cui i 6 centri avrebbero introdotto le novità. Alla fine tutti i centri hanno ricevuto la formazione ma secondo un ordine casuale e non in base a criteri di tipo amministrativo; ciò ha consentito di realizzare una valutazione affidabile degli effetti del nuovo servizio sui tassi di recidiva*.

* I risultati finali debbono ancora essere pubblicati. Per maggiori dettagli sul disegno dello studio si veda Pearson *et al*. (2010).

valutazione degli effetti che i servizi di disintossicazione dall'alcool offerti ai detenuti generano sulla propensione a ricommettere reati, si potrebbero misurare: il numero degli invii al servizio, l'effettiva partecipazione degli alcolisti ai corsi, l'assunzione di alcool (rilevata con un questionario) o il tasso di recidiva dei detenuti.

In questo caso la recidiva è il risultato che ci interesserebbe di più; tali dati potrebbero però essere difficili da raccogliere e il beneficio potrebbe richiedere anni per diventare evidente. Per questo motivo, si potrebbe prendere in considerazione un esito surrogato (*proxy*) misurando l'effettiva partecipazione al servizio di assistenza. In alternativa, si potrebbero misurare sia la frequenza del servizio, in grado di dare risultati sul processo, sia il compimento di reati nei 24 mesi successivi, come risultato di lungo periodo (*follow-up*). Gli invii dei detenuti al servizio da parte degli ufficiali giudiziari sarebbero i dati più facili da misurare; anche se molto immediata, tale misura non è però molto utile se pensiamo che la grandezza più idonea a misurare il successo dell'intervento sia la recidiva (un approfondimento si trova nel riquadro 13). Per identificare la misura di risultato più idonea è spesso molto utile consultare sia i tecnici (che conoscono le tecniche dello studio), sia i *policy maker* (cui sono note disponibilità, tempistiche e metodologie di raccolta dei dati amministrativi).

Fase 3 - Decidere l'unità di randomizzazione

Dopo aver deciso quale risultato misurare (Fase 2), dobbiamo decidere chi o che cosa randomizzare: l'unità di randomizzazione.

L'unità di randomizzazione è, nella maggior parte dei casi, costituita da singole persone, per esempio assegnate in modo casuale a ricevere uno o un altro trattamento

Riquadro 13 - Argomenti a favore (e contro) l'utilizzo di esiti surrogati

Un esito surrogato è un'approssimazione del vero e proprio risultato atteso; per esempio, i tassi di nuove condanne (*reconviction*) sono utilizzati come surrogato dei tassi di recidiva, perché sono molto più facili da misurare (in quanto le persone potrebbero anche non essere mai arrestate per i crimini commessi). L'utilizzo di un esito surrogato è tanto più accettabile quanto più vi è evidenza che esso sia un buon predittore del vero risultato d'interesse. Purtroppo, le autodichiarazioni sui cambiamenti di comportamento, per quanto facili da rilevare, sono un indice assai approssimativo dei cambiamenti effettivi. I rispondenti tendono, infatti, a fornire risposte distorte, alterando i valori dichiarati nella direzione attesa dall'intervistatore (distorsione da desiderabilità sociale). Ad esempio, i partecipanti a un programma per il recupero della forma fisica potrebbero essere portati a esagerare la quantità dichiarata di esercizio fisico svolto durante gli allenamenti.

Se si rende necessario utilizzare esiti surrogati perché i risultati finali diverranno disponibili solo a lungo termine, vale comunque sempre la pena di aspettare anche questi in modo da poter confermare (o meno) i risultati intermedi. La letteratura riporta numerosi esempi, in medicina clinica, nei quali le prove iniziali, basate su esiti surrogati erano fuorvianti. Ad esempio, il trattamento dell'osteoporosi con i fluoruri fu creduto efficace perché aumentava la densità ossea, indicatore tradizionalmente utilizzato come esito surrogato. Tuttavia, è stato successivamente dimostrato che questo tipo di trattamento determina anche un aumento di alcuni tipi di fratture, un risultato che i pazienti osteoporotici preferiscono evitare accuratamente (Riggs *et al.*, 1990; Rothwell, 2005).

medico o uno o un altro programma educativo. Tuttavia, l'unità di randomizzazione può anche essere costituita da un gruppo di persone che ruotano attorno a un ente, soprattutto se l'intervento è offerto contestualmente a più persone. Ad esempio, intere scuole potrebbero essere selezionate in modo casuale per utilizzare un nuovo metodo d'insegnamento o continuare con quello attuale; alcuni Centri per l'impiego, selezionati casualmente, potrebbero offrire un nuovo programma di formazione. Infine, l'unità di randomizzazione potrebbe essere costituita anche da un'intera area geografica: ad esempio, alcune autorità locali potrebbero essere selezionate casualmente per fornire uno dei due nuovi programmi di prevenzione sanitaria o diversi metodi di riciclaggio dei rifiuti (riquadro 14).

Alla fine della sperimentazione, i risultati possono essere misurati sia sui singoli individui sia sull'intera unità di randomizzazione, scegliendo tra precisione e praticità. Ad esempio, anche se a intere classi potrebbero essere attribuiti casualmente diversi metodi d'insegnamento, utilizzare per la valutazione i risultati degli apprendimenti dei singoli studenti permette una maggiore precisione delle stime.

La questione se utilizzare individui, istituzioni o aree come unità di randomizzazione dipende solitamente da considerazioni pratiche. Negli studi clinici, per esempio, è generalmente possibile fornire il placebo o il farmaco da testare direttamente agli individui. Nel caso delle politiche pubbliche ciò non è sempre possibile. Di seguito si trovano due esempi di modi diversi con i quali il *Behavioural Insights Team* ha deciso che tipo di unità di randomizzazione utilizzare:

Riquadro 14 - Sfruttare le varianti locali della politica

Le autorità locali sono luoghi ideali per testare sul campo nuove politiche pubbliche. Possono collaborare con altre autorità locali per sperimentare politiche diverse o implementare interventi diversi in alcuni quartieri o aree zone estratte a caso; le autorità locali possono quindi utilizzare proficuamente gli studi randomizzati per valutare quali politiche funzionano meglio.

Un esempio di questo approccio è la sperimentazione condotta dall'autorità locale del Nord Trafford (Regno Unito) per confrontare metodi diversi di promozione del riciclaggio dei rifiuti. Le unità di randomizzazione in questo caso erano intere strade. Metà delle strade scelte per la sperimentazione sono state assegnate in modo casuale all'intervento di promozione del riciclaggio con una campagna porta a porta. Le famiglie del gruppo sperimentale hanno mostrato tassi di riciclaggio più alti rispetto alle 3.000 famiglie del gruppo di controllo.

La crescita del tasso di riciclaggio è stata del 5% nel breve periodo a fronte di un costo della campagna porta a porta di circa £24 per ogni famiglia che ha deciso di iniziare a riciclare i rifiuti (Cotterill *et al.*, 2009; John *et al.*, 2011). Sulla base di queste informazioni, l'autorità locale ha potuto stabilire che la riduzione dei costi di discarica prodotta dalla campagna porta a porta giustifica la sua estensione.

- Individuale: Quando si sperimentano messaggi diversi nelle lettere fiscali, è ovviamente possibile inviare diverse lettere a persone diverse, per cui l'unità di randomizzazione sono i singoli debitori.

- Istituzionale: Quando si esegue una valutazione del sostegno dei Centri per l'impiego all'avvio al lavoro, non è possibile assegnare casualmente diversi interventi alle persone in cerca di lavoro, per cui l'unità randomizzazione sono i team dei Centri per l'impiego (consulenti che aiutano i disoccupati a trovare lavoro).

Come già visto in passaggi precedenti, anche in questo caso sarà utile discutere dell'unità di randomizzazione con un esperto. Sarà anche importante considerare varie implicazioni della scelta dell'unità di randomizzazione. Tale scelta produce conseguenze rilevanti sul numero di persone che occorre coinvolgere nell'esperimento: considerare come unità di randomizzazione istituzioni o aree geografiche comporta, quasi sempre, la necessità di un campione più ampio di individui e l'adozione di particolari metodi di analisi.

Ci possono poi essere anche altri elementi da considerare: ad esempio, in una valutazione di un programma che incentiva la frequenza ai corsi per l'educazione degli adulti, i ricercatori hanno scelto di randomizzare intere classi, anche se sarebbe stato possibile randomizzare singoli partecipanti (riquadro 15). Tale soluzione è stata adottata per evitare che nella stessa classe si trovassero allievi che avrebbero goduto dell'incentivo e allievi che ne rimanessero esclusi. Ciò avrebbe, infatti, potuto influenzare negativamente la partecipazione ai corsi e sarebbe stato poi impossibile distinguere l'effetto indesiderato indotto da questo problema da quello generato dal programma. Inoltre, è di fondamentale importanza che gli individui siano reclutati per lo studio prima della randomizzazione, altrimenti la robustezza dell'esperimento ne risulterebbe compromessa. Per esempio, se le persone che gestiscono un esperimento conoscono a quale gruppo sarà assegnato un potenziale partecipante prima che questo sia formalmente reclutato nello studio, questo potrebbe influenzare la

loro decisione di coinvolgerlo. Un ricercatore o un operatore che crede sinceramente nel nuovo intervento potrebbe scegliere - forse anche inconsciamente - di non inserire partecipanti ritenuti "senza speranza" nel gruppo d'intervento. Ciò però renderebbe non rappresentativi i partecipanti di ciascun gruppo "determinato casualmente". Questo tipo di problema può essere evitato semplicemente reclutando formalmente i partecipanti nell'esperimento prima della loro randomizzazione.

Riquadro 15 - Quando l'unità randomizzazione dovrebbe essere costituita da gruppi e non da singoli individui

I parassiti intestinali (*anchilostoma*) infettano quasi un quarto della popolazione mondiale, soprattutto nei paesi in via di sviluppo. Si tratta di una causa molto comune di assenza a scuola e alcuni ricercatori statunitensi hanno collaborato con il Ministero della Salute degli Stati Uniti per valutare se un programma di trattamento dei parassiti intestinali rivolto ai bambini sarebbe in grado ridurre l'assenteismo scolastico.

Per rispondere a questa domanda è stato quindi realizzato un esperimento randomizzato controllato nel quale intere scuole sono state sottoposte a un trattamento vermifugo di massa, mentre altre hanno continuato, come al solito, senza alcun trattamento. In questo caso, l'utilizzo di una randomizzazione individuale sarebbe stata inappropriata, perché la compresenza di bambini – trattati e non - nelle stesse scuole avrebbe ridotto artificialmente la probabilità che il gruppo di controllo contraesse l'infezione, semplicemente per il venir meno del rischio di contagio rappresentato dai propri compagni beneficiari della cura.

Settantacinque scuole elementari in zone rurali del Kenya sono quindi state coinvolte in uno studio che ha dimostrato che il programma di trattamento vermifugo riduce l'assenteismo di un quarto (Miguel, Kremer, 2004). L'aumento della frequenza scolastica è risultato particolarmente marcato nei bambini più piccoli. Lo studio ha anche dimostrato che il programma ha anche un ottimo rapporto costo/efficacia. La deverminazione è infatti in grado di produrre un allungamento del periodo di scolarizzazione a costi molto bassi: un anno aggiuntivo di scolarizzazione prodotto attraverso un intervento di questo tipo costerebbe $3,50 dollari per ciascuno studente, a fronte di costi superiori ai $100 di programmi in grado di produrre analoghi risultati (uniformi scolastiche gratuite e simili)°.

° A questo proposito si vedano ancora Karlan e Appel (2011), cit.

Fase 4 - Determinare la numerosità necessaria a ottenere risultati affidabili

Per trarre conclusioni di *policy* da uno studio randomizzato controllato, questo deve essere condotto su un campione di dimensioni sufficienti. Se la dimensione del campione è abbastanza grande, si può essere ragionevolmente certi di potere attribuire gli effetti rilevati all'intervento e non al caso. Se abbiamo deciso che l'unità randomizzazione sarà costituita da istituzioni o da aree (riquadro 16), è molto probabile che avremo bisogno di coinvolgere nell'esperimento un maggior numero di persone di quelle che sarebbero state necessarie se avessimo deciso di randomizzare dei singoli. Alcuni semplici e preliminari calcoli di potenza consentiranno di determinare quante unità (individui, istituzioni etc.) dovranno essere incluse nei gruppi d'intervento e di controllo del programma. Si consiglia quindi di lavorare con esperti che possano vantare un'effettiva esperienza negli studi randomizzati in modo da assicurare che questi calcoli siano svolti correttamente.

Se il programma che s'intende valutare produce un grande vantaggio (effetto), questo potrà essere rilevato anche con una dimensione relativamente piccola del campione. Rilevare differenze (effetti) più piccole tra gli interventi richiederà invece un numero maggiore di partecipanti, è quindi molto importante essere cauti sin dall'inizio circa le possibilità di successo di un intervento. Molti interventi - se non la maggior parte - producono infatti effetti relativamente piccoli. Un esempio di quanti partecipanti sia necessario includere in un esperimento: assegnare 800 persone a due gruppi composti ciascuno di circa 400 persone, dovrebbe darci circa 8 possibilità su 10 di vedere una differenza del 10%, sempre che tale differenza esista veramente.

Per esempio, immaginiamo che il governo voglia incoraggiare la gente a votare e verificare l'efficacia dell'invio di brevi messaggi di testo (SMS) come promemoria agli elettori nella mattina del giorno delle elezioni. A questo scopo sono scelti 800 elettori da osservare: 400 nel gruppo di controllo che non riceverà alcun messaggio e 400 nel gruppo sperimentale, che riceverà gli SMS. Se l'affluenza sarà del 50% nel gruppo di controllo, con un campione di queste dimensioni avremmo l'80% di probabilità di vedere una variazione dal 50% al 60% (10 punti percentuali di variazione). Se invece volessimo essere in grado di rilevare una differenza più piccola, avremmo bisogno di un campione di dimensioni più grandi.

Occorre comunque tenere in considerazione il costo di reclutamento di ogni persona in più e l'impatto (dimensione dell'effetto e potenziali risparmi sui costi) dell'intervento che viene misurato. Talvolta, rilevare anche una differenza modesta è molto utile, in particolare, se l'intervento è poco o per nulla costoso. Ad esempio, se stiamo cambiando lo stile o il contenuto di una lettera per incoraggiare il tempestivo pagamento delle imposte, il costo aggiuntivo sarà molto piccolo, i costi di spedizione saranno sostenuti comunque e la raccolta dei dati di risultato (in questo caso, le date di pagamento) è già organizzata. Al contrario, se volessimo aumentare la percentuale degli iscritti a un programma di assistenza alla ricerca del lavoro (es. la *Job Seekers' Allowance*) offrendo loro un servizio individuale di consulenza, i costi sarebbero molto rilevanti e quindi dovremmo aspettarci di vedere un effetto proporzionalmente più grande per giustificare l'esecuzione dell'esperimento. Tuttavia, anche per gli interventi costosi, se a impatti piccoli (in termini di dimensione dell'effetto) corrispondessero ritorni potenzialmente grandi in termini di risparmio (ad esempio la riduzione del numero delle persone che richiedono prestazioni), l'argomento potrebbe essere assai convincente per decidere lo svolgimento di un esperimento randomizzato controllato.

Fase 5 – Selezionare i beneficiari con un metodo di randomizzazione robusto

Come abbiamo visto, l'assegnazione casuale delle unità di studio ai gruppi di trattamento e di controllo è la caratteristica che rende lo studio randomizzato controllato il migliore metodo disponibile per la valutazione delle politiche: esso ci permette infatti di essere certi che i due gruppi siano equivalenti rispetto a tutti i fattori chiave. Nel caso degli interventi nel settore dell'istruzione, ad esempio, questi potrebbero includere lo *status* socio-economico, il genere e il titolo di studio degli studenti. Ci sono diversi modi con i quali fattori distorsivi possono insinuarsi nel processo di randomizzazione, è quindi molto importante assicurarsi che questo sia impostato correttamente, fin dall'inizio, per evitare problemi nelle fasi successive.

Ci sono molti casi che mostrano come chi ha un interesse personale nell'esito di uno studio possa tentare di allocare le persone in modo non casuale, anche se inconsciamente. Ad esempio, se un esperimento relativo a una politica di reinserimento al

lavoro assegna le persone ai gruppi di intervento e di controllo sulla base del numero della loro tessera di previdenza sociale - per cui ai numeri dispari è assegnato il nuovo intervento - una persona addetta al reclutamento potrebbe escludere, consciamente o inconsciamente, alcune persone con tessera di numero dispari dall'esperimento nel caso sospettasse che questi non si comporterebbero bene, in modo da proteggere l'intervento e i suoi risultati. Un comportamento del genere introduce distorsioni nell'esperimento, per cui il metodo di randomizzazione deve poter resistere a interferenze di questo tipo. Ci sono molte organizzazioni indipendenti come, ad esempio, unità di studi clinici, che possono aiutare a costruire un sistema di randomizzazione sicuro per evitare problemi di cattiva assegnazione. Un sistema classico è un'assegnazione ai gruppi mediante un generatore di numeri casuali delle persone già formalmente reclutate nella sperimentazione (per i motivi già visti sopra).

Al momento della randomizzazione, possono anche essere adottate ulteriori misure per garantire l'equilibrio dei gruppi rispetto alle caratteristiche delle persone che li compongono: ad esempio, assicurare la stessa composizione per età e genere di ciascun gruppo. Ciò è particolarmente importante negli studi più piccoli che hanno minore potenza statistica.

Riquadro 16 - Costruire variazioni negli interventi per consentire i *test*

La sperimentazione comporta il confronto degli effetti di un intervento (una possibile nuova politica) con quelli di un altro (ad esempio, la politica attuale). Un test appropriato richiede ovviamente che le variazioni della politica (ad esempio nuova e attuale) possano essere realizzate contemporaneamente. In alcuni casi questo è abbastanza semplice - alcune scuole potrebbero continuare a servire i soliti pasti scolastici, mentre altre potrebbero fornirli secondo i nuovi standard nutrizionali - e l'effetto sui comportamenti potrebbe essere misurato. In altri casi, l'organizzazione dei servizi esistenti potrebbe rendere difficile offrire interventi differenti nello stesso momento.

Ad esempio, anche se un ente locale desiderasse testare l'efficacia di un modulo di richiesta semplificato, il suo sistema di corrispondenza potrebbe essere esternalizzato e/o non essere in grado di stampare più di un modello di lettera. Per questo motivo, è consigliabile che nello sviluppare nuovi sistemi o sottoscrivere nuovi contratti di servizio con i fornitori, i *policy maker* si assicurino la possibilità di introdurre alcune variazioni nel corso del tempo. Anche se questo può comportare alcuni costi iniziali, la possibilità di testare diverse versioni dei servizi potrebbe giustificare tali spese. Proprio per questo motivo, il Ministero del lavoro britannico prevede specificamente che i sistemi informatici attraverso i quali è gestita l'erogazione di alcune prestazioni di *welfare* (es. Universal Credit) abbiano la possibilità di fornire versioni differenti del servizio, in modo da garantire la possibilità di svolgere sperimentazioni e scoprire ciò che funziona e, nel caso, estenderne l'applicazione.



Fase 6 - Proporre gli interventi ai diversi gruppi

Una volta che gli individui, istituzioni o le aree geografiche sono state assegnate casualmente al gruppo sperimentale o di controllo, viene il momento di somministrare l'intervento. Si pensi, ad esempio, all'introduzione di un nuovo programma educativo in un determinato gruppo di scuole mentre il servizio resta invariato nelle altre. Quando il *Behavioural Insights Team* ha effettuato l'esperimento per verificare gli effetti degli SMS sulla propensione a pagare le multe, per esempio, gli individui nei gruppi d'intervento hanno ricevuto diversi tipi di messaggio, mentre quelli del gruppo di controllo non ne hanno ricevuto nessuno.

Una questione molto importante in questa fase è disporre di un sistema per monitorare l'intervento in modo da garantire che questo sia somministrato esattamente come previsto dal progetto (protocollo). Nel caso dell'invio dell'SMS, il monitoraggio è servito a garantire che i messaggi giusti arrivassero effettivamente alle persone selezionate.

L'uso della valutazione di processo per monitorare la corretta implementazione dell'intervento consente di ottenere risultati significativi e risolvere tempestivamente eventuali problemi gestionali. In particolare, è molto importante per garantire che l'intervento oggetto di valutazione rifletta effettivamente quello che - in caso di successo - sarà portato a regime. Facendo ancora riferimento all'esperimento degli SMS, è emerso che i numeri di cellulare non sono sempre disponibili per tutte le persone coinvolte. Se, durante l'esperimento, si fossero spesi tempo e denaro per controllare e correggere i numeri di telefono errati, i risultati sarebbero stati sicuramente migliori, ma non avrebbero tenuto conto di quanto sarebbe probabilmente successo nelle reali condizioni di esercizio.

3.2 Impara

Fase 7 - Misurare i risultati e l'impatto degli interventi

Dopo l'avvio dell'intervento è necessario misurare i risultati. Tempistica e metodo di valutazione dovrebbero essere già stati decisi prima della randomizzazione. Questi dipenderanno da quanto rapidamente l'intervento produrrà i suoi effetti, elemento

che sarà diverso per ogni intervento. L'invio dei messaggi per incoraggiare le persone a pagare le proprie multe ha bisogno solo di alcune settimane di applicazione per produrre i propri effetti, mentre per un intervento che modifichi il *curriculum* scolastico potrebbero essere necessari anche diversi anni.

Oltre che per la valutazione del risultato principale, può essere utile anche raccogliere dati sul processo d'implementazione. Ad esempio, nella valutazione di differenti misure alternative al carcere, si potrebbero raccogliere dati presso le diverse agenzie cui i beneficiari dell'intervento sono stati inviati, in modo da aiutare a interpretare i risultati. In questo caso, una riduzione della recidiva potrebbe, ad esempio, corrispondere a un aumento degli invii ai corsi per la gestione della collera. Questi risultati secondari non potranno essere interpretati con la stessa certezza di quelli principali ma potranno servire per sviluppare nuove ipotesi da sottoporre a ulteriori prove (si veda il riquadro 17).

Molti esperimenti prevedono anche la raccolta di dati qualitativi utili a interpretare i risultati, sostenere la futura attuazione della politica e fungere da guida per ulteriori ricerche e/o miglioramenti dell'intervento. Ciò non è sempre strettamente necessario ma, se si prevede di svolgere una ricerca qualitativa, l'ideale è farla coincidere con l'esperimento, osservando gli stessi partecipanti sui quali saranno anche già disponibili numerose informazioni.

Riquadro 17 - Utilizzo intelligente dei dati

Siamo spesso interessati a sapere se una politica è effettivamente efficace per un campione rappresentativo della popolazione generale. In alcuni casi, tuttavia, potremmo essere interessati a scoprire se alcuni gruppi (ad esempio uomini e donne, oppure giovani e anziani) rispondono in modo diverso dagli altri. È importante decidere in via preliminare se siamo interessati a segmentare il campione in questo modo; condurre un'analisi su sottogruppi dopo che i dati sono già stati raccolti, comporta forti rischi di riduzione della potenza statistica e della validità interna ed esterna. Tuttavia, se all'interno di un sottogruppo dovesse evidenziarsi un'inaspettata tendenza (per esempio gli uomini potrebbero risultare più sensibili rispetto alle donne ai promemoria degli appuntamenti per effettuare visite specialistiche), si potrebbe svolgere un apposito approfondimento successivo per verificare che tale risultato sia effettivamente robusto. Di solito è comunque utile la raccolta di dati aggiuntivi (ad esempio età, sesso) che serviranno a segmentare il campione e supportare la ricerca futura.

Come abbiamo visto, a volte, l'esperimento può far emergere risultati imprevisti. Ad esempio, si potrebbero notare grandi fluttuazioni nel tempo dell'efficacia di un incentivo a sostegno degli interventi di coibentazione dei sottotetti e scoprire che ciò è correlato alle variazioni di temperatura. Questo fenomeno potrebbe suggerire che le persone siano più ricettive a rispondere a questo tipo d'incentivi quando il clima è più freddo. In quanto frutto di un'analisi non pianificata, i risultati potrebbero non potersi considerarsi definitivi; tuttavia, nessuna informazione deve andare sprecata e questo risultato potrebbe rivelarsi prezioso per orientare la ricerca futura.

3.3 Adatta

Fase 8 - Adattare l'intervento tenendo conto dei risultati della sperimentazione

Mettere a regime interventi che producono effetti positivi è spesso più facile che cancellare politiche che si sono rivelate inefficaci. Qualsiasi esperimento svolto dovrebbe essere comunque considerato un successo, a prescindere dal risultato prodotto. Un esperimento che non mostra effetti o, addirittura, produce effetti negativi è prezioso quanto quello che mostra la generazione dei benefici desiderati.

Ad esempio, l'esperimento condotto dal Ministero del lavoro britannico sugli *Incapacity benefit* - già citato nel riquadro 3 - ha mostrato che l'intervento proposto non produceva gli effetti attesi. Tuttavia, se possiamo essere sicuri che l'esperimento è stato condotto correttamente - ha regolarmente identificato (prima dell'effettuazione del test) la misura di efficacia, ha coinvolto un campione così ampio da rilevare un effetto statisticamente significativo sulle variabili di interesse (anche queste identificate prima di iniziare) - le informazioni ricavate sono comunque utili e consentono di imparare molto dall'esperienza.

Quando gli interventi si dimostrano inefficaci, si deve prendere in considerazione un "disinvestimento razionale" che porti alla chiusura dell'intervento. In tali casi il denaro risparmiato può essere speso ad esempio per progettare, valutare e realizzare interventi efficaci nello stesso campo (ad esempio gli *Incapacity benefit*).

Quando un esperimento randomizzato controllato è stato completato è buona pratica pubblicarne i risultati, con ricchezza di informazioni sui metodi utilizzati in modo che altri possano valutarne la correttezza. È anche importante descrivere compiutamente l'intervento e i suoi beneficiari, in modo che altri possano replicarlo al meglio (riquadro 18).

Un documento utile che può guidare la stesura del rapporto di valutazione degli studi randomizzati è il *CONSORT statement*[10], utilizzato negli studi medici e, sempre più spesso, anche in esperimenti non medici. Le linee guida CONSORT garantiscono che le parti fondamentali dell'esperimento e degli interventi siano accuratamente descritte in modo da consentirne la massima replicabilità.

Idealmente, il protocollo della sperimentazione dovrebbe essere pubblicato prima che l'esperimento abbia inizio, cosicché le persone possano, per tempo, criticare e proporre miglioramenti. La pubblicazione anticipata del protocollo rende immediatamente chiaro se il risultato atteso riportato nel rapporto di ricerca coincide con quello dichiarato prima dell'inizio dell'esperimento.

Fase 9 - Ritornare alla fase 1 per continuare a migliorare

L'o studio randomizzato andrebbe visto non tanto come uno strumento per valutare un singolo programma in un determinato momento, ma come una parte del continuo processo d'innovazione e miglioramento delle politiche. La replica dei risultati di un esperimento è particolarmente importante se l'intervento è offerto a un diverso

10 http://www.consort-statement.org/consort-statement

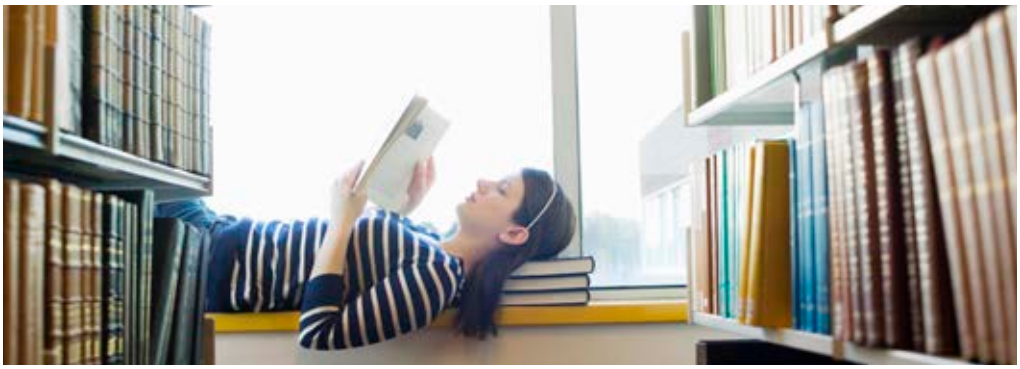
segmento di popolazione rispetto a quello dell'esperimento originale. È anche utile poter lavorare sui risultati dell'esperimento in modo da identificare nuovi metodi per migliorare ulteriormente l'efficacia, specie quando l'esperimento mirava proprio a identificare le componenti con il maggiore impatto. In un lavoro recente con HMRC (l'Agenzia delle Entrate e delle Dogane del Regno Unito), il Behavioural Insights Team ha cercato di identificare i messaggi più efficaci per aiutare le persone a rispettare norme e scadenze fiscali. Ciò ha consentito di trarre diversi insegnamenti su quello ciò che sembra funzionare meglio - per esempio, utilizzare contenuti e formati semplici e informare i debitori che la maggior parte dei residenti nelle loro zone ha già pagato le proprie imposte.

Tuttavia, piuttosto che conservare le lezioni apprese pensando di aver raggiunto la perfezione, è assai più utile continuare a pensare alla possibilità di ulteriori affinamenti: ci sono, per esempio, altri modi per semplificare le forme e aiutare i contribuenti a rispettare le norme o ci sono messaggi che potrebbero rivelarsi più efficaci su tipologie specifiche di contribuenti? Lo stesso pensiero può essere applicato a tutti i settori delle politiche pubbliche: dal miglioramento della *performance* scolastica, all'assistenza nella ricerca di occupazione.

Il miglioramento continuo, in questo senso, è l'ultimo ma forse il più importante aspetto della metodologia 'sperimenta, impara, adatta', che presuppone che non siamo mai in grado di sapere in anticipo quanto potremo migliorare qualsiasi politica pubblica.

Riquadro 18 - La riduzione della mortalità dei pazienti nelle case di cura

Le vaccinazioni antinfluenzali sono regolarmente offerte a tutti i gruppi a rischio, tra cui gli anziani, all'avvicinarsi della stagione influenzale. Nelle case di cura, tuttavia, il virus influenzale può anche essere introdotto dal personale. Nel 2003 è stato condotto un esperimento randomizzato controllato in modo da determinare se il costo di una campagna per la vaccinazione del personale avrebbe: a) aumentato i tassi di vaccinazione del personale e b) avuto effetti positivi sulla salute dei pazienti. Oltre 40 case di cura sono state assegnate in modo casuale ai gruppi di controllo (nessuna modifica rispetto allo status quo) o sperimentale (che prevedeva la realizzazione di una campagna di sensibilizzazione del personale cui era anche offerta la possibilità di vaccinarsi). Nel corso di due stagioni influenzali, l'assunzione dei vaccini è risultata - non sorprendentemente - significativamente più alta nelle case di cura incluse nel gruppo sperimentale. Ancora più importante, la mortalità per qualsiasi causa degli ospiti è risultata inferiore di 5 decessi ogni 100 ospiti (Hayward, *et al.* 2006). Questa ricerca ha contribuito alla promozione di una campagna nazionale per la vaccinazione del personale delle case di cura e altre campagne, anche internazionali, a sostegno della vaccinazione degli operatori sanitari.



BIBLIOGRAFIA

Banerjee A.V., Cole S., Duflo E., Linden L. (2007), Remedying education: Evidence from two randomised experiments in India. *Quarterly Journal of Economics*, 122, 3: 1235-1264.

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

Brooks G., Burton M., Cole P., Miles J., Torgerson C., Torgerson D. (2008), Randomised controlled trial of incentives to improve attendance at adult literacy classes. *Oxford Review of Education*, 34, 5: 493-504.

Christensen C. (2003), *The Innovator's Dilemma: The revolutionary book that will change the way you do business.* HarperBusiness: New York.

Cotterill S., John P., Liu H., Nomura H. (2009), *How to get those recycling boxes out: A randomised controlled trial of a door to door recycling service*. Manchester: IPEG.

Davenport T.H., Harris J.G. (2007), *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Review Press.

Delta Airlines Magazine (2007), 0915, 22 - www.deltaskymag.com.

DWP – Department of Work and Pensions (2006), *Research Report 382. Jobseekers Allowance intervention pilots quantitative evaluation*. London: UK, DWP, Research & Statistics – http://research.dwp.gov.uk.

DWP – Department of Work and Pensions (2006), *Research Report 342. Impacts of the Job Retention and Rehabilitation Pilot*. London: UK, DWP, Research & Statistics – http://research.dwp.gov.uk.

Edwards P. (ed.) (2005), Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury – outcomes at 6 months. *Lancet*, 365, 9475: 1957-1959.

Finckenauer J.O. (1982) *Scared Straight and the Panacea Phenomenon*. Englewood Cliffs, NJ: Prentice-Hall.

Harford T. (2011), *Adapt: Why success always starts with failure*. London: Little, Brown.

BIBLIOGRAFIA

Hayward A. (ed.) (2006), Effectiveness of influenza vaccine programme for care home staff to prevent death, morbidity and health service use among residents; cluster randomised control trial. *British Medical Journal*, 333, 7581: 1241-1247.

John P., Cotterill S., Richardson L., Moseley A., Smith G., Stoker G, Wales C. (2011), *Nudge, nudge, think, think: Using experiments to change civic behaviour*. London: Bloomsbury Academic.

Karlan D., Appel J. (2011), *More than good intentions: How a new economics is helping to solve global poverty*. New York: Dutton.

Luca M. (2011), Reviews, reputation, and revenue: The case of Yelp.com. Boston: *Harvard Business School Working Paper* n. 12-016.

MacMillan H.L. (ed.) (2009), Interventions to prevent child maltreatment and associated impairment. *Lancet*, 363, 9659: 250-266.

Miguel E., Kremer M. (2004), Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72, 1: 159-217.

Pearson D., Torgerson D., McDougall C., Bowles R. (2010), A parable of two agencies, one of which randomises. *Annals of the American Academy of Political & Social Sciences*, 628, 1: 11-29.

Petrosino A., Turpin-Petrosino C., Buehler J. (2003), *Scared Straight and other juvenile awareness programs for preventing juvenile delinquency. A Systematic Review. Campbell Systematic Reviews* – http://www.campbellcollaboration.org.

Riggs B.L., Hodgson S.F., O' Fallon W.M. (1990), Effect of fluoride treatment on fracture rate in postmenopausal women with osteoporosis. *New England Journal of Medicine*, 322, 12: 802–809.

Rothwell P.M. (2005), External validity of randomised controlled trials:"To whom do the results of this trial apply?" *Lancet*, 365, 9453: 82-93.

Shepherd J. (2007), The production and management of evidence for public service reform. *Evidence and Policy,* 3, 2: 231-251.

Taleb N.N. (2007), *The Black Swan: The impact of the highly improbable*. London: Allen Lane.

The Social Research Unit (2012), *Youth justice: Cost and benefits. Investing in Children, 2.1*. DSRU – Dartington Social Research Unit - http://investinginchildren.eu/.

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

Quaderno N.20 #BICIttadini - Interventi a favore della mobilità ciclistica

Quaderno N.21 Resilienza tra territorio e comunità - Approcci, strategie, temi e casi

Quaderno N.22 Favorire la coesione sociale con le biblioteche. Valutazione del bando

Quaderno N.23 Il "mercato" dei lasciti testamentari. Nuove stime per Italia e Lombardia (2014-2030)

Quaderno N.24 Il bando abitare sociale temporaneo. Mappatura e analisi dei progetti finanziati (2000-2013)

Quaderno N.25 Lo sviluppo dei Green Jobs. Uno scenario di evoluzione quantitativa e qualitativa e alcune ipotesi di adeguamento dei percorsi formativi

Quaderno N.26 House rich, cash poor. Come rendere liquida la ricchezza rappresentata dalla casa di abitazione

Quaderno N.27 Bando materiali avanzati 2003-2013 – Progetti e risultati

Quaderno N.28 Sperimenta, impara, adatta. Sviluppare politiche pubbliche con gli esperimenti randomizzati controllati

doi: 10.4460/2018quaderno28

fondazione
cariplo