La sfida della misurazione degli effetti causali di un intervento*

Enrico Rettore
Università degli Studi di Trento
e FBK-IRVAPP

L'articolo presenta in modo non tecnico la logica e i principali metodi per la valutazione degli effetti di un intervento. L'esito controfattuale è il *benchmark* rispetto al quale si compara ciò che si osserva in presenza dell'esposizione. Essendo tale esito non osservabile, occorre trovare un gruppo di soggetti non esposti all'intervento – il gruppo di confronto – che lo approssimi in modo appropriato. Dopo aver presentato sinteticamente alcuni studi di caso, l'articolo si conclude sottolineando l'importanza di predisporre per tempo le condizioni per la valutazione, identificando il gruppo di confronto in anticipo rispetto all'inizio di un intervento e pianificando i modi e i tempi di reperimento delle informazioni necessarie per la valutazione.

This article introduces, in a non-technical way, the logic and the main methods for evaluating the effect of an intervention. The counterfactual event is the benchmark against which to compare what happens in the presence of the intervention. The counterfactual event being unobservable, a group of subjects unexposed to the intervention needs to be identified – the comparison group – to suitably approximate it. After presenting some case studies the article concludes by stressing the relevance of designing the evaluation in advance, identifying the comparison group and planning the operations for collecting all the relevant information.

DOI: 10.1485/SINAPPSI_2018/3_429

Citazione
Rettore E. (2018), La sfida della misurazione degli effetti causali di un intervento, *Sinappsi*, VIII, n.3, pp.6-19

Parole chiave
Esiti fattuali e controfattuali
Disegni sperimentali e osservazionali
Valutazioni prospettiche e retrospettive

Key words
Factual and counterfactual outcomes
Experimental and observational designs
Prospective and retrospective evaluations

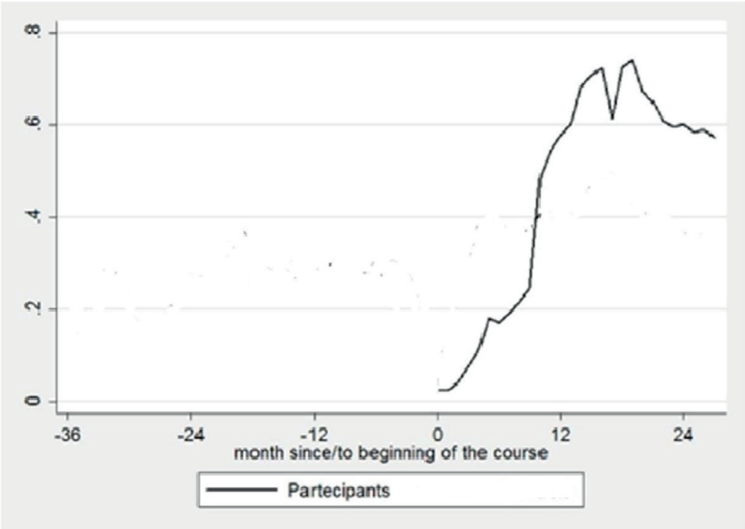
1. Effetto causale di un intervento: di cosa si tratta

Un intervento produce degli effetti se *fa* capitare qualcosa. Cioè se quel qualcosa *non* capiterebbe in assenza dell'intervento stesso. Per fissare le idee, consideriamo il caso dei corsi di formazione riservati a giovani disoccupati, nella fascia di età 20-29, diplomati della scuola media superiore, attivati nella Provincia autonoma di Trento negli anni 2010 e 2011. La figura 1 riporta i tassi di occupazione dei partecipanti a tali corsi nei mesi a partire dall'inizio dei corsi stessi fino al 27° successivo. Questi corsi hanno avuto un effetto sul tasso di occupazione dei partecipanti se gli *stessi* soggetti, negli *stessi* mesi, in assenza dei corsi, avrebbero sperimentato un tasso di occupazione *diverso* da quello osservato.

L'esempio chiarisce da subito la distinzione, fondamentale per il seguito, tra ricerca degli effetti di una causa e ricerca delle cause di un effetto. Applicata all'esempio appena introdotto, la logica e i metodi discus-

* Basato sulla relazione tenuta al *2018 Meeting of the Community of Practice on Counterfactual Impact Evaluation of ESF interventions*, Atene, 14 giugno 2018.

Figura 1
Tassi di occupazione dei partecipanti ai corsi di formazione nei mesi dall'inizio dei corsi stessi fino al 27° successivo, Provincia autonoma di Trento 2010 e 2011



Fonte: elaborazioni dell'Autore

si in questo articolo servono a rispondere al quesito se uno *specifico evento* – la partecipazione ai corsi di formazione – abbia o meno avuto un *effetto* sui tassi di occupazione dei partecipanti. Non rispondono al quesito – molto più generale – di quali siano gli *eventi* che influiscono sulla *probabilità* di lavorare.

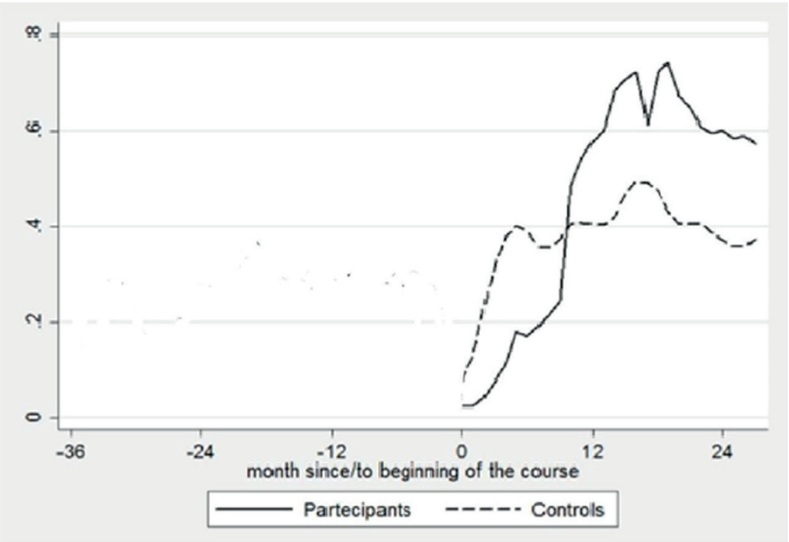
Questa è la trasparente definizione di effetto causale di un intervento, basata sul confronto *ipotetico* tra ciò che succede ai soggetti esposti all'intervento – il loro risultato *fattuale* – e ciò che succederebbe agli stessi soggetti nello stesso periodo temporale se non venissero esposti all'intervento – il loro risultato *controfattuale*.

Questa definizione trasparente di effetto causale, basata sul confronto tra ciò che osserviamo in presenza dell'intervento e ciò che avremmo osservato se l'intervento non avesse avuto luogo, pone palesemente un gigantesco problema quando si passa alla pratica: il risultato fattuale è osservabile, il risultato controfattuale non lo è. Proseguendo l'esempio precedente, in figura 1 manca un ingrediente essenziale per la misura dell'effetto causale dei corsi: il risultato controfattuale dei partecipanti ai corsi. Come misuriamo il tasso di occupazione che i partecipanti ai corsi avrebbero sperimentato se non avessero partecipato ai corsi?

L'unica via di uscita praticabile è rappresentata dal ricorso a una *approssimazione*. Cioè, la sostituzione del risultato controfattuale dei soggetti esposti all'intervento con il risultato fattuale di un *opportuno* gruppo di soggetti non esposti, il *gruppo di confronto*. Continuando l'esempio precedente, in figura 2 compaiono anche i tassi di occupazione di un gruppo di soggetti – stessa fascia di età, livello di istruzione, area di residenza – non partecipanti ai corsi.

Ma come stabilire se questo gruppo di non partecipanti approssima in modo accurato il tasso controfattuale dei partecipanti? La verifica diretta è ovviamente impossibile, essendo il risultato controfattuale dei partecipanti… controfattuale. Come si deve procedere in pratica per selezionare l'*opportuno* gruppo di confronto?

Figura 2
Tassi di occupazione dei partecipanti e dei non partecipanti ai corsi di formazione nei mesi dall'inizio dei corsi stessi al 27° successivo, Provincia autonoma di Trento 2010 e 2011



Fonte: elaborazioni dell'Autore

Questo è il passaggio cruciale di *qualsiasi* valutazione dell'impatto di un intervento. Più in generale, è il problema che si pone *sempre* quando si vogliono identificare relazioni di causa ed effetto.

Qui appare in tutta la sua rilevanza la distinzione introdotta poco sopra tra ricerca degli effetti di una causa e ricerca delle cause di un effetto. Per l'identificazione degli effetti di *uno* specifico evento serve trovare due gruppi di soggetti il più possibile simili tra loro, gli uni che abbiano sperimentato *quell'evento*, gli altri che non l'abbiano sperimentato. Un compito impegnativo, ma in molte circostanze praticabile, come vedremo nei paragrafi successivi. L'identificazione dell'insieme delle cause di un effetto moltiplica per il numero delle possibili cause lo sforzo richiesto per identificare l'effetto di una singola causa; richiede cioè di replicare *separatamente per ogni (ipotetica) causa* la ricerca dei gruppi di soggetti comparabili, gli uni esposti, gli altri non esposti a quella (ipotetica) causa. Un compito arduo nelle circostanze tipiche delle scienze sociali.

La soluzione del problema di reperimento del gruppo di confronto dipende dal grado di controllo che il valutatore ha sul *processo di selezione*, ovvero il processo in accordo al quale alcuni soggetti risultano esposti all'intervento, altri no. Il valutatore può *decidere* chi verrà esposto all'intervento e chi no? Se la risposta è affermativa – vedremo oltre in quali situazioni è plausibile lo sia – la soluzione consiste nella scelta *casuale* dei soggetti destinati, e dei soggetti non destinati, all'esposizione all'intervento. A tutti gli effetti, l'estrazione di una lotteria (par. 2).

Se al contrario il valutatore non ha alcun grado di controllo sul processo di selezione – oppure se lo ha solo parzialmente – l'obiettivo non cambia: si tratta di selezionare un gruppo di soggetti non esposti all'intervento che consentano di approssimare accuratamente il risultato medio controfattuale dei soggetti esposti (par. 3).

2. Il *Randomized Control Trial* (RCT)

In un RCT la scelta tra i soggetti coinvolti nello studio di coloro che verranno esposti all'intervento avviene in modo *casuale*, ad esempio mediante il lancio di una moneta. Nella nostra lingua, ne risulta un gioco di parole: per scoprire l'effetto *causale* dell'intervento si ricorre a una scelta *casuale* dei soggetti esposti all'intervento.

Se il numero di soggetti inclusi nei due gruppi è *sufficientemente elevato*, la scelta casuale assicura che i due gruppi siano *mediamente* equivalenti in ogni aspetto. La conseguenza fondamentale di tale equivalenza è che se entrambi i gruppi *non* fossero esposti all'intervento risulterebbero *mediamente* equivalenti anche rispetto agli esiti finali. Vale a dire che l'esito fattuale – osservato – dei soggetti non esposti è *mediamente* eguale all'esito controfattuale – non osservato – dei soggetti esposti. Ne consegue che se i due gruppi sperimentano *in media* esiti diversi, tali differenze sono attribuibili univocamente all'*unica* differenza tra i due gruppi: gli uni sono stati esposti all'intervento, gli altri no.

Alcune precisazioni sono a questo punto indispensabili. Entrambi i gruppi – gli esposti e i non esposti all'intervento – devono essere *sufficientemente numerosi*. La scelta casuale, infatti, dà luogo a due gruppi che per effetto del caso possono risultare diversi. L'unico modo per mantenere entro margini accettabili tali differenze dovute al caso è disporre di un adeguato numero di soggetti in entrambi i gruppi. Sono disponibili semplici regole statistiche per determinare la numerosità dei due gruppi (si veda ad esempio Bloom 2006).

I due gruppi selezionati in modo casuale sono equivalenti *in media*: ne deriva che il confronto tra gli esiti medi ottenuti dai due gruppi dà luogo a una stima dell'*effetto medio* dell'intervento. Altrimenti detto, non c'è modo di scoprire l'effetto dell'intervento su ogni singolo soggetto che vi viene esposto. Ci sono soggetti che beneficiano maggiormente di altri dall'esposizione all'intervento? Ci sono soggetti che vengono addirittura danneggiati dall'esposizione? Si tratta, palesemente, di domande della massima importanza. Parziali risposte sono possibili svolgendo l'analisi separatamente per vari sottogruppi definiti secondo caratteristiche individuali (età, genere, istruzione…). Il prezzo da pagare per questo affinamento dei risultati, in ragione di quanto detto appena sopra, è che la numerosità del gruppo degli esposti e del gruppo di confronto deve essere sufficientemente elevata, distintamente per ognuno dei sottogruppi presi in considerazione.

Gli esperimenti con assegnazione casualizzata sono stati utilizzati per valutare l'impatto di innumerevoli interventi pubblici, sia in Paesi sviluppati che in Paesi in via di sviluppo (si veda ad esempio la rassegna in Duflo e Banerjee 2017). Sono particolarmente adatti nel caso dei cosiddetti *interventi pilota*, vale a dire interventi messi in atto su piccola scala, con il preciso scopo di stabilire *se* sono efficaci.

La principale obiezione che viene mossa dai critici dell'utilizzo di RCT nelle scienze sociali è di natura etica: la lotteria esclude ingiustamente dall'intervento soggetti che potrebbero beneficiarne. Ma almeno nel caso degli interventi pilota si tratta di una obiezione infondata, proprio perché un intervento pilota viene messo in atto perché *non si sa* se l'intervento consenta di ottenere risultati desiderabili.

Quando si passa dalla teoria alla pratica sorgono inevitabilmente problemi. Nel caso di RCT il problema di gran lunga più frequente è il mancato rispetto dell'assegnazione a uno dei due gruppi di una frazione dei soggetti coinvolti nello studio (*non compliance* nella letteratura internazionale): da un lato, soggetti assegnati all'esposizione all'intervento che si sottraggono (*no shows*); dall'altro, soggetti assegnati al gruppo di confronto che vogliono, e trovano il modo di, essere esposti all'intervento (*cross-overs*).

Il problema deriva dal fatto che, avendo a che fare con esseri umani, si deve inevitabilmente tenere conto del fatto che hanno desideri, preferenze, idiosincrasie e che almeno entro certi limiti sono liberi di scegliere. Per questo motivo, meglio pensare a RCT come uno strumento utile a migliorare i *disegni di valutazione osservazionali*, tema oggetto del prossimo paragrafo.

3. Cosa si può fare se non è possibile condurre un RCT? I disegni osservazionali

Che il valutatore abbia o meno pieno controllo sul processo di selezione fa la differenza per il *modo* in cui può procedere alla selezione del gruppo di confronto, ma è del tutto irrilevante per il *fine* della selezione. L'obiettivo è quello anticipato nel paragrafo precedente: la selezione di un gruppo di soggetti non esposti all'intervento che consentano di approssimare accuratamente il risultato medio controfattuale dei soggetti esposti.

La differenza importante rispetto al caso di RCT è data dal fatto che il valutatore non ha alcun controllo sul processo di selezione (o lo ha solo parzialmente). Può solo osservare (o parzialmente influire su) ciò che accade attorno a lui. Una ovvia pre-condizione perché la valutazione dell'impatto abbia luogo è che oltre ai soggetti esposti all'intervento ci siano anche soggetti che non vi sono esposti. Per esemplificare le quattro principali strategie osservazionali a disposizione del valutatore, nel seguito si considerano alcuni studi di caso.

a) Il controllo delle differenze osservabili tra esposti e non esposti all'intervento: il matching *(e i metodi connessi)*

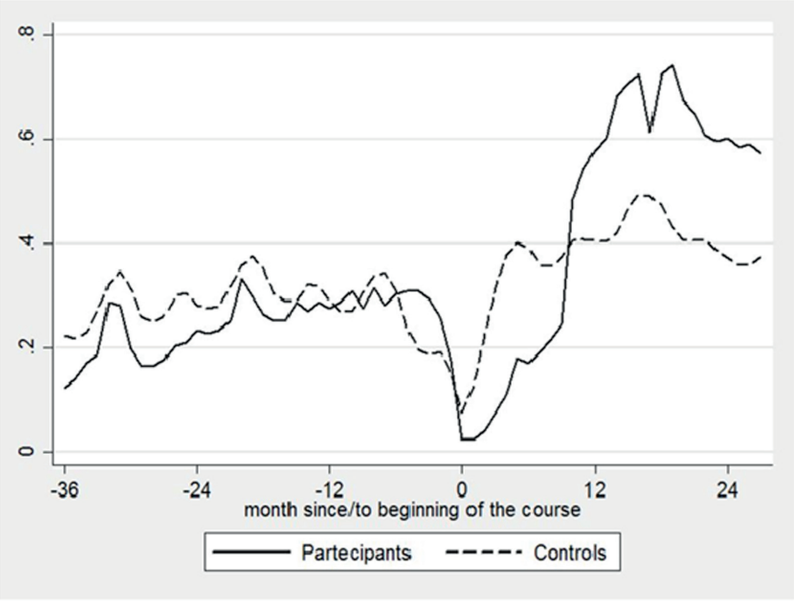
Nell'esempio considerato in questo paragrafo i soggetti ammissibili ai corsi di formazione *scelgono* liberamente se partecipare o meno. Palesemente, si tratta di un caso nel quale il valutatore non ha la minima possibilità di influire sul processo di selezione. Nel caso specifico, il gruppo di confronto è costituito da soggetti nella stessa fascia di età, di pari livello di istruzione, residenti nella stessa area geografica e disoccupati alla data di inizio dei corsi. Si tratta di soggetti che hanno avuto l'opportunità di partecipare a uno dei corsi e che hanno *scelto* di non partecipare.

Il problema che il valutatore deve risolvere in questo caso è dato dalla possibile *non comparabilità* dei due gruppi: non essendo stati selezionati casualmente, non è detto che il risultato osservato sui non partecipanti approssimi in modo accurato ciò che si sarebbe osservato sui partecipanti se non avessero partecipato ai corsi. Un modo semplice per valutare la comparabilità dei due gruppi è dato dall'analisi della loro storia occupazionale *precedente* l'inizio dei corsi. Per essere effettivamente comparabili, i due gruppi, nei mesi precedenti l'inizio dei corsi – vale a dire nei mesi durante i quali l'effetto dei corsi non può ancora manifestarsi – dovrebbero presentare tassi di occupazione comparabili.

Come si vede in figura 3, così non è nel caso considerato: ci sono chiari segni che i partecipanti ai corsi presentano tassi di occupazione sistematicamente inferiori ai non partecipanti, nell'arco dei tre anni precedenti l'inizio dei corsi. Questa evidenza è sufficiente per dubitare che il gruppo dei soggetti non partecipanti selezionato in questo modo fornisca una buona approssimazione del risultato controfattuale dei soggetti esposti.

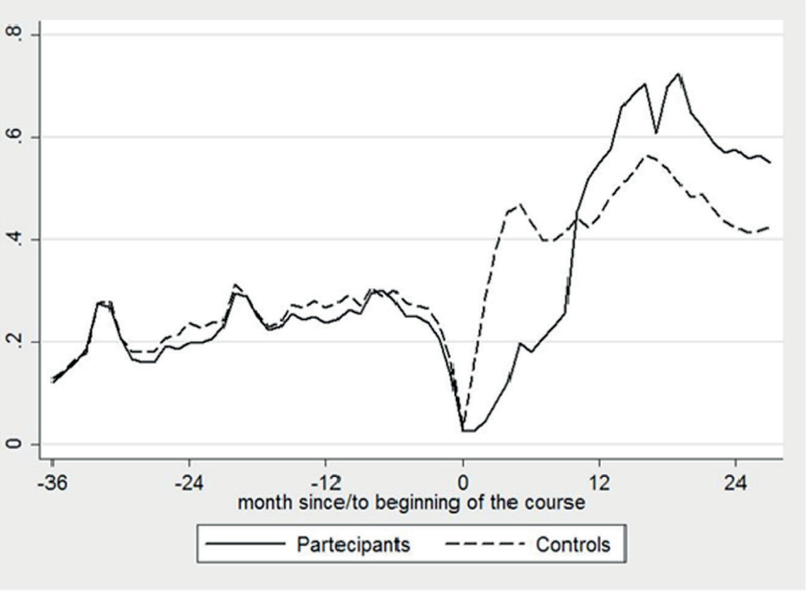
Una soluzione praticabile consiste nel *mimare ex post* – per quanto possibile – ciò che si farebbe se fosse possibile condurre un RCT. Operativamente, si tratta di selezionare tra i soggetti non partecipanti *solo* quelli che nei periodi precedenti l'inizio dei corsi presentano caratteristiche comparabili a quelle dei partecipanti. Nel caso considerato, per ogni partecipante è disponibile una dettagliata descrizione delle sue esperienze lavorative nei tre anni precedenti l'inizio dei corsi (oltre a informazioni socio-demografiche). Il

Figura 3
Tassi di occupazione dei partecipanti e dei non partecipanti ai corsi di formazione da 36 mesi prima a 27 mesi dopo l'inizio dei corsi stessi, Provincia autonoma di Trento 2010 e 2011



Fonte: elaborazioni dell'Autore

Figura 4
Tassi di occupazione dei partecipanti e dei non partecipanti ai corsi di formazione, abbinati mediante *matching*, da 36 mesi prima a 27 mesi dopo l'inizio dei corsi stessi, Provincia autonoma di Trento 2010 e 2011



Fonte: elaborazioni dell'Autore

gruppo di confronto viene selezionato scegliendo tra i non partecipanti coloro che risultano più simili ai partecipanti rispetto all'insieme di queste informazioni.

Sono disponibili varie tecniche statistiche per implementare questa idea: *matching*, regressione, ponderazione (per i dettagli tecnici si veda ad esempio Angrist e Pischke 2014).

In figura 4 sono presentati i risultati dell'analisi. Nei 36 mesi precedenti l'inizio dei corsi i tassi di occupazione dei due gruppi sono – per costruzione! – pressoché identici: il gruppo di confronto è stato selezionato includendovi *solo* i soggetti con carriere occupazionali comparabili a quelle dei partecipanti ai corsi. Nei primi mesi dopo l'inizio dei corsi il tasso di occupazione dei partecipanti è *inferiore* a quello dei non partecipanti. Si tratta del cosiddetto effetto *lock-in*, particolarmente pronunciato attorno al 6° mese: nei mesi di frequenza dei corsi, i corsisti sono impegnati in un'attività che rende loro difficile – se non impossibile – lo svolgimento di un lavoro. A partire dal 10° mese, il tasso di occupazione dei partecipanti risulta sistematicamente superiore a quello del gruppo di confronto di 10-15 punti percentuali.

La differenza fondamentale tra un RCT e la soluzione appena descritta – che mima ex-post un RCT – consiste nel fatto che i due gruppi risultanti da una scelta casuale sono equivalenti in media rispetto a *qualsiasi* caratteristica non modificabile dall'intervento, precedente e successiva alla messa in atto dell'intervento stesso. Mentre la soluzione che mima ex-post un RCT dà luogo a due gruppi equivalenti in media rispetto alle sole caratteristiche in base alle quali i due gruppi vengono

forzati a essere equivalenti, vale a dire caratteristiche osservate *prima* che l'intervento abbia luogo.

I metodi basati sul controllo delle caratteristiche osservabili danno il loro meglio quando è disponibile un ampio insieme di informazioni relative ai periodi precedenti la messa in atto dell'intervento, su tutti i soggetti inclusi nello studio. È il caso dello studio appena presentato, basato su informazioni relative alle carriere occupazionali tratte da archivi amministrativi.

b) I disegni basati su una discontinuità nella regola di assegnazione

Il caso qui considerato si riferisce agli effetti della regolarizzazione degli immigrati sul loro tasso di criminalità. I dati si riferiscono a coloro che hanno presentato domanda di regolarizzazione a dicembre 2007. All'epoca vigeva una regola basata sul meccanismo delle quote (stabilite innanzitutto per anno e per provincia, con vari altri dettagli; si veda Pinotti 2017). A partire dalle ore 8.00 del giorno stabilito, i candidati datori di lavoro dei richiedenti la regolarizzazione hanno presentato la domanda online. Le regolarizzazioni sono state concesse secondo l'ordine di presentazione delle domande, fino all'esaurimento della quota stabilita.

L'autore dello studio ha ottenuto l'accesso ai dati amministrativi relativi ai crimini commessi da *tutti* gli immigrati che hanno presentato domanda di regolarizzazione a dicembre 2007, sia regolarizzati che non, con l'intento di misurare l'effetto della regolarizzazione sulla probabilità di commettere un crimine.

Palesemente, si tratta di un processo di selezione dei regolarizzati/non regolarizzati basato su una *regola sistematica* – il momento della presentazione della domanda – anche in questo caso completamente diverso da quello di un RCT. Per cui, il confronto diretto tra il tasso di criminalità, rispettivamente, dei regolarizzati e dei non regolarizzati *non* identifica l'effetto causale cercato, per il solito motivo, secondo cui il risultato medio osservato sui non regolarizzati potrebbe non corrispondere al risultato medio controfattuale dei regolarizzati.

In questa situazione, una buona soluzione – anche se non priva di controindicazioni, come vedremo nel seguito – consiste nel restringere l'analisi al sottoinsieme dei soggetti che hanno ottenuto, oppure mancato, la regolarizzazione *di poco*. Cioè coloro per i quali una *marginale* differenza nel momento di presentazione della domanda avrebbe invertito l'esito della selezione.

Qui sta la differenza chiave rispetto al caso trattato nel punto precedente (a). In questo caso l'ottenimento della regolarizzazione *non* risulta dalla scelta dei soggetti aventi diritto. È determinata *solo* dal momento della presentazione della domanda. Ne consegue che i primi richiedenti esclusi sono *molto simili* agli ultimi tra i richiedenti ammessi, rispetto all'*unica* caratteristica individuale rilevante per l'ammissione alla regolarizzazione: il momento della presentazione della domanda. Cioè, i primi tra i richiedenti esclusi forniscono una buona approssimazione del risultato controfattuale degli ultimi tra i richiedenti ammessi (e viceversa).

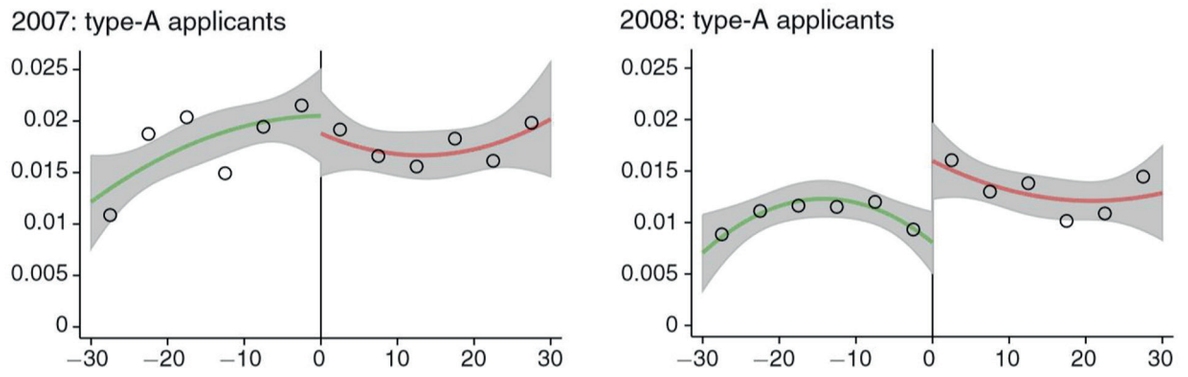
Questo confronto al margine – tra i primi esclusi e gli ultimi ammessi – consente di misurare l'effetto della regolarizzazione sugli *ultimi tra gli ammessi* (o, equivalentemente, sui *primi tra i non ammessi*). In figura 5 sono rappresentati i tassi di criminalità (asse verticale) di coloro che hanno presentato domanda di regolarizzazione, al variare del momento di presentazione della domanda (asse orizzontale), rispettivamente, per l'anno successivo e per l'anno precedente la presentazione della domanda. Convenzionalmente, lo zero lungo l'asse orizzontale corrisponde al momento di presentazione della domanda dell'ultimo ammesso; a sinistra dello zero si osservano i tassi di criminalità degli ammessi; a destra, dei non ammessi.

Nell'anno successivo alla presentazione della domanda – il 2008 – gli ultimi tra gli ammessi – i soggetti appena a sinistra del tempo zero – hanno compiuto un numero medio di crimini circa pari a 0,008, mentre i primi tra gli esclusi – i soggetti appena a destra del tempo zero – hanno un numero medio di crimini pari a 0,015, circa il doppio del valore osservato per gli ammessi. Vale a dire che in questo caso, almeno nel breve termine, la regolarizzazione ha causato una rilevante diminuzione del tasso di criminalità.

A conferma del fatto che i primi esclusi forniscono una buona approssimazione del risultato controfattuale degli ultimi ammessi, il grafico a sinistra di figura 5 mostra che nell'anno precedente la presentazione della domanda – vale a dire l'anno nel quale *nessuno* tra i soggetti considerati godeva dei benefici della regolarizzazione – non si osserva alcuna differenza nel tasso di criminalità tra i due gruppi, rispettivamente, appena sopra e appena sotto lo zero.

L'effetto della regolarizzazione misurato in que-

Figura 5
Numero medio di crimini commessi per richiedente nell'anno precedente (a sinistra) e nell'anno successivo (a destra) la presentazione della domanda secondo l'esito della domanda*



* In verde i regolarizzati, in rosso i non regolarizzati. L'asse orizzontale misura il tempo in minuti dal momento della domanda dell'ultimo ammesso.

Fonte: Pinotti, 2017

sto modo si riferisce, palesemente, ai soli ammessi al margine. Per quanto detto sulla somiglianza tra soggetti, rispettivamente, appena sopra e appena sotto il momento nel quale si è esaurita la quota, questo effetto ci dice anche quanto avrebbero guadagnato dalla regolarizzazione i primi tra gli esclusi.

Ma qual è l'effetto della regolarizzazione sugli ammessi che hanno presentato la domanda con largo anticipo sull'ultimo momento utile, ad esempio 30 minuti prima? Quanto avrebbe influito la regolarizzazione sui non ammessi che hanno presentato la domanda con forte ritardo sull'ultimo momento utile, ad esempio 30 minuti dopo?

La strategia di stima dell'effetto basata sul confronto attorno alla soglia *non* consente di rispondere a queste domande. Vale a dire che questa strategia consente di ottenere una stima credibile dell'effetto dell'intervento al prezzo – alto – di restringere l'analisi a un particolare sottoinsieme di soggetti. Gli studi più recenti stanno sviluppando metodi per generalizzare la stima ottenuta per i soggetti al margine a una popolazione più ampia (si vedano Battistin e Rettore 2008; Angrist e Rokkanen 2015).

Nella letteratura, questo problema è noto con il nome di mancato supporto comune (*common support*, nella letteratura anglosassone). Il problema ha luogo quando il valutatore cerca soggetti non esposti all'intervento che siano comparabili ai soggetti esposti, e almeno per qualche soggetto esposto la ricerca fallisce; ne deriva che il confronto tra soggetti esposti e non esposti risulta praticabile solo per un sottoinsieme dei soggetti esposti. Nel caso appena esaminato – un caso estremo di mancanza di supporto comune – ad esempio, per gli ammessi che hanno presentato domanda 30 minuti prima dell'ultimo momento utile non è possibile ottenere una stima dell'effetto perché non c'è modo di ottenere soggetti non ammessi che siano comparabili rispetto al momento di presentazione della domanda: coloro che hanno presentato domanda 30 minuti prima del momento utile sono stati *tutti* ammessi. Per cui, il confronto tra ammessi e non ammessi risulta praticabile solo per i soggetti a ridosso dell'ultimo momento utile.

La differenza con RCT è palese. Almeno sulla carta (ma si veda la discussione del successivo punto *d)*), in un RCT l'*intero* insieme dei soggetti esposti risulta mediamente comparabile con l'insieme dei soggetti non esposti. Vale a dire che in un RCT *non* si pone alcun problema di mancato supporto comune.

c) Il metodo della differenza di differenze (Diff-in-Diff's)

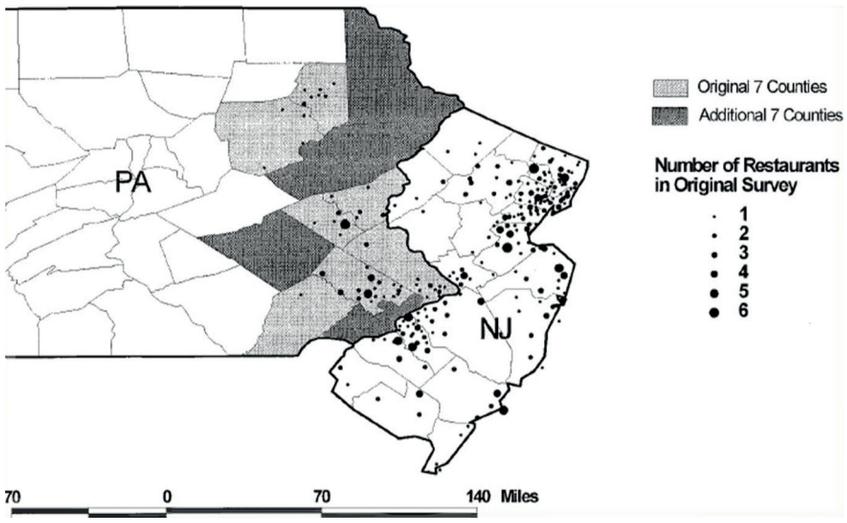
Il caso qui considerato è tratto da un vecchio articolo (Card e Krueger 1994), ma tratta di un tema sul quale è tuttora in corso un dibattito acceso: quali sono gli effetti dell'introduzione di un salario minimo di legge?

Con decorrenza 1° aprile 1991, una legge federale degli Stati Uniti ha alzato il salario orario minimo da $3.35 a $4.25. Nel New Jersey il salario minimo è stato

ulteriormente alzato a $5.05 a partire dal 1 aprile 1992. All'epoca, si trattava del salario minimo più elevato tra gli Stati dell'Unione.

Gli autori dello studio hanno condotto un'indagine su un campione di ristoranti fast food (Burger King, KFC, Wendy's, Roy Rogers) operanti in New Jersey e nell'area limitrofa della Pennsylvania. La prima intervista è stata condotta tra febbraio e marzo del 1992, vale a dire poco prima che diventasse operativo il nuovo salario minimo in New Jersey. A quella data, in *entrambi* gli Stati vigeva un salario minimo pari a $4.25. Sono state rilevate, tra le altre, informazioni sul numero di dipendenti, sul loro salario, sul prezzo dei prodotti venduti.

Figura 6
La mappa geografica del disegno di valutazione



Fonte: Card e Krueger, 1994

Tra novembre e dicembre dello stesso anno i ristoranti inclusi nel campione sono stati ricontattati aggiornando le informazioni già rilevate a inizio anno. A quella data, il salario minimo della Pennsylvania era immutato, mentre in New Jersey era salito a $5.05 per effetto della riforma.

I due gruppi di ristoranti sono geograficamente prossimi (si veda la figura 6). Ciò nonostante, vi sono alcune differenze degne di nota già *prima* della riforma del New Jersey, cioè in un periodo nel quale il salario minimo era esattamente lo stesso nei due Stati. In tabella 1 si osserva in particolare che i ristoranti della Pennsylvania (PA) presentano un numero medio di dipendenti superiore e hanno una maggiore percentuale di dipendenti part-time.

Questa evidenza è sufficiente per dubitare che i ristoranti della Pennsylvania approssimino in modo accurato il risultato controfattuale dei ristoranti del New Jersey.

A novembre 1992, i dati rilevati nella seconda intervista svolta presso i ristoranti mostrano innanzitutto che la riforma del New Jersey ha prodotto gli effetti attesi sulla distribuzione dei salari orari (si veda la figura 7): a febbraio/marzo la distribuzione del salario orario è approssimativamente la stessa nei due Stati; a fine anno, nel New Jersey sono spariti i salari orari inferiori al nuovo minimo di legge.

Tabella 1
I risultati dell'indagine precedente l'aumento del salario minimo in New Jersey

	NJ	PA	t stat.
Means in Wave 1			
a. FTE employment	20.4 (0.51)	23.3 (1.35)	-2.0
b. Percentage full-time employees	32.8 (1.3)	35.0 (2.7)	-0.7
c. Starting wage	4.61 (0.02)	4.63 (0.04)	-0.4
d. Wage = $4.25 (percentage)	30.5 (2.5)	32.9 (5.3)	-0.4

Fonte: Card e Krueger, 1994

In tabella 2 sono riportati i dati relativi al numero medio di dipendenti in entrambi gli Stati e in entrambe le occasioni di indagine[1]. Si osserva che tra

1 Si tratta del numero di dipendenti equivalenti a tempo pieno: un dipendente part-time conta quanto mezzo dipendente a tempo pieno.

Figura 7
La distribuzione dei salari orari in Pennsylvania e New Jersey a febbraio e a novembre 1992

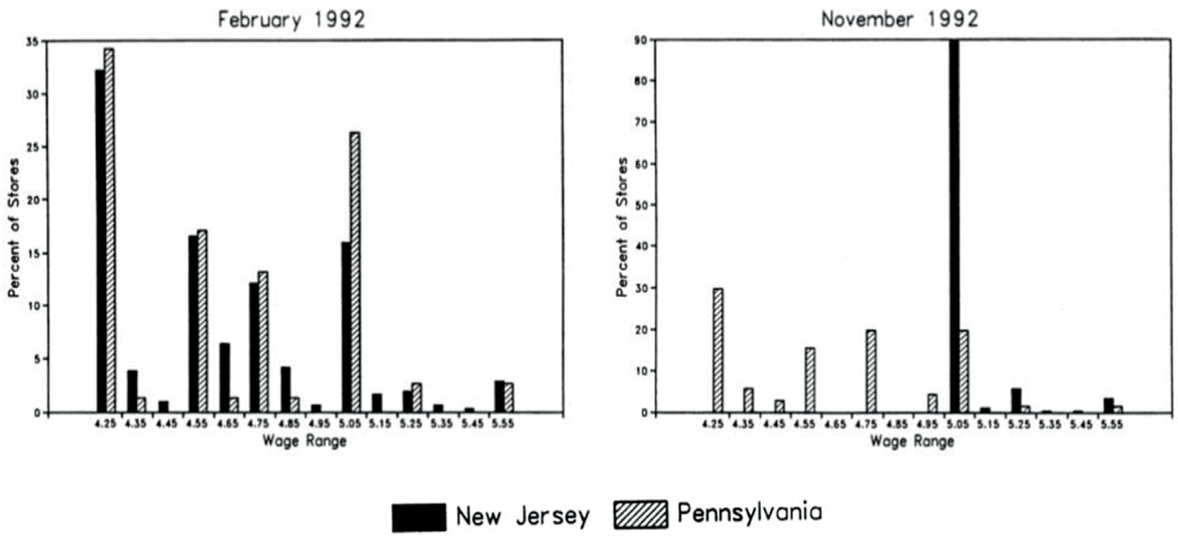


FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

Fonte: Card e Krueger, 1994

marzo e novembre 1992 il numero medio di dipendenti per ristorante è *calato* in Pennsylvania di oltre due unità, mentre è rimasto pressoché *stabile* in New Jersey.

La soluzione adottata dagli autori per stimare l'effetto della riforma del salario minimo in New Jersey poggia su un'ipotesi detta di *trend comune*. L'ipotesi afferma che in assenza della riforma, la *variazione* dell'occupazione per i ristoranti del New Jersey sarebbe stata identica a quella osservata per i ristoranti della Pennsylvania. Questa ipotesi stabilisce un'equivalenza tra la *variazione controfattuale* dell'occupazione in New Jersey e la corrispondente *variazione fattuale* in Pennsylvania. Si tratta di un'ipotesi del tutto plausibile dato che la recessione che ha colpito la Pennsylvania (e tutta la costa Est degli US) nel 1992 è facile pensare che avrebbe causato una riduzione dell'occupazione in New Jersey del tutto analoga se in questo Stato non fosse intervenuta una variazione del salario minimo.

Ed è un'ipotesi – questo è il punto centrale per la valutazione dell'impatto – che consente di stimare quale *sarebbe stato* il numero medio di occupati nei ristoranti del New Jersey in assenza della riforma del salario minimo:

occupazione *controfattuale* in NJ in nov. 1992 =
FTE in NJ in mar. 1992 +
(FTE in PA in nov. 1992 - FTE in PA in mar. 1992)

In sintesi, se non fosse avvenuta la riforma del salario minimo, a novembre 1992 l'occupazione in New Jersey sarebbe stata pari all'occupazione lì osservata a marzo, aumentata di una quantità pari alla variazione dell'occupazione osservata in Pennsylvania nello stes-

Tabella 2
Numero medio di dipendenti (equivalenti al tempo pieno) in Pennsylvania e New Jersey prima (febbraio 1992) e dopo (novembre 1992) l'aumento del salario minimo in New Jersey*

	PA	NJ	NJ-PA
FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

* (Standard errors in parentheses)

Fonte: Card e Krueger, 1994

so arco di tempo. La differenza tra l'occupazione fattuale a novembre 1992 e l'occupazione controfattuale calcolata nel modo detto fornisce una stima dell'impatto della riforma del salario minimo.

Si ricava immediatamente che tale stima è data dalla cosiddetta *differenza di differenze* (*Diff-in-Diff's*):

Impatto = (FTE in NJ in nov. 1992 - FTE in NJ in mar. 1992) –
(FTE in PA in nov. 1992 - FTE in PA in mar. 1992)

Cioè, la differenza tra la variazione dell'occupazione osservata in New Jersey e la corrispondente variazione osservata in Pennsylvania. Nel caso considerato, ne risulta un effetto *positivo* dell'aumento del salario minimo sull'occupazione media dei ristoranti fast food. Questo risultato, inaspettato alla luce di quanto suggerisce la teoria economica ortodossa, ha suscitato un acceso dibattito (si veda ad esempio Card e Krueger 2000)[2].

Anche in questo caso, la differenza con RCT è palese. La validità del metodo *Diff-in-Diff's* poggia in modo cruciale sulla correttezza dell'ipotesi di trend comune: in assenza dell'intervento oggetto di valutazione, i due gruppi di unità – nel caso qui discusso, i due gruppi di ristoranti – si muoverebbero lungo due traiettorie parallele. In un RCT, in assenza dell'intervento è certo – per il modo in cui vengono selezionati! – che i due gruppi di unità si muoverebbero lungo due traiettorie *coincidenti*, non solo parallele.

L'ipotesi di trend comune non può, ovviamente, essere sottoposta a verifica, per il solito motivo: il trend controfattuale del gruppo esposto all'intervento non è osservabile. Tuttavia, un test indiretto consiste nel comparare le traiettorie dei due gruppi nei periodi *precedenti* la messa in atto dell'intervento. Scoprire che le due traiettorie *non* sono parallele smentirebbe l'ipotesi di trend comune togliendo credibilità a questa strategia di stima dell'impatto.

d) Il metodo delle variabili strumentali

Il caso qui discusso è tratto da un recente articolo (Martini *et al*. 2018). Illustra in modo trasparente che, anche quando sulla carta è il valutatore a decidere chi deve essere esposto all'intervento, nella pratica succede che i suoi sforzi per tenere sotto controllo il processo di selezione siano almeno in parte vanificati dal comportamento sul campo dei vari attori coinvolti. Ne consegue che uno studio nato per essere RCT finisce per essere in buona misura osservazionale.

Lavoro&Psiche è un intervento pilota messo in atto per studiare gli effetti di una misura di sostegno alle persone affette da disturbi mentali nella ricerca di un lavoro. La componente di gran lunga più importante dell'intervento consiste nella disponibilità di un *job coach*, interamente dedicato al supporto nella ricerca di lavoro di un piccolo numero di soggetti (non più di 12-13 per addetto).

Nel corso del 2010 sono stati reclutati per lo studio 311 soggetti, bipartiti *casualmente* tra gruppo di trattamento e gruppo di controllo. Il periodo sperimentale ha avuto luogo tra l'inizio del 2011 e la fine del 2012. In questo periodo i soggetti inclusi nel gruppo di trattamento sono stati seguiti dal *job coach* previsto dall'intervento. Gli effetti dell'intervento sono stati valutati con riferimento agli esiti occupazionali nel corso del 2013.

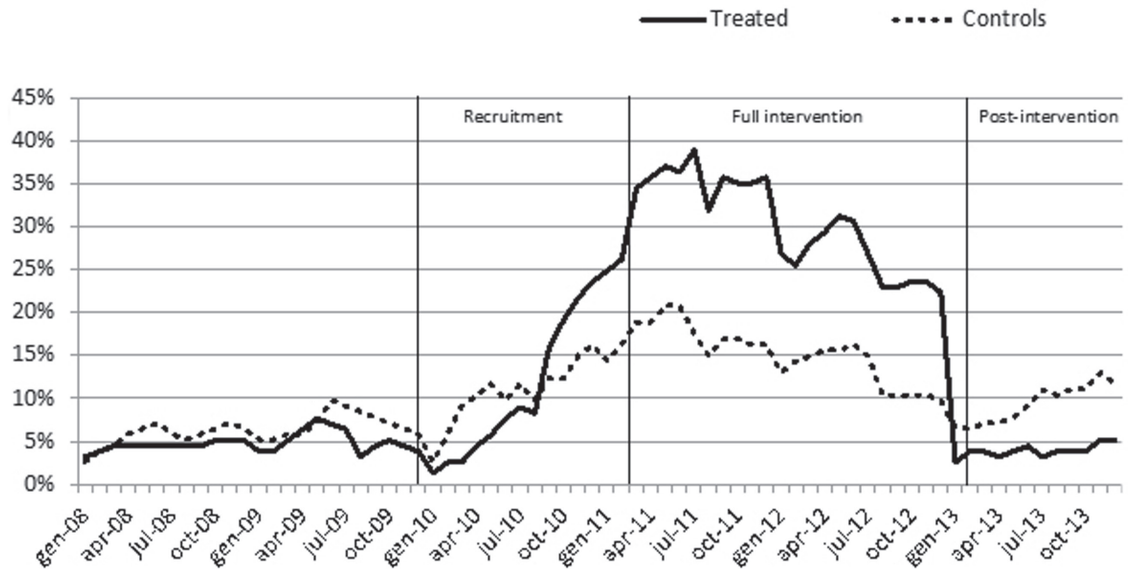
Il dato di fatto emerso dall'analisi di implementazione dell'intervento è che il principale ruolo svolto dal *job coach* durante il biennio dell'esperimento è consistito nel facilitare l'accesso a un *tirocinio* ai soggetti assistiti, il che equivale a dire che l'effetto causale dell'intervento sugli esiti lavorativi successivi al biennio sperimentale – se c'è – passa dall'esperienza di un tirocinio.

La figura 8 mostra l'andamento nel tempo della partecipazione a un tirocinio separatamente per il gruppo di trattamento e per il gruppo di controllo. Nel caso del gruppo di trattamento è evidente il ruolo svolto dal *job coach*: il tasso di partecipazione cresce in modo vistoso all'inizio del periodo sperimentale, si mantiene 20-30 punti percentuali sopra il corrispondente tasso relativo al gruppo di controllo e cala in modo altrettanto vistoso alla fine del periodo stesso.

Ma il punto critico per la valutazione è dato dalla parziale violazione dell'assegnazione al trattamento: non tutti i soggetti inclusi nel gruppo di trattamento sperimentano un periodo di tirocinio, mentre lo spe-

2 Peraltro, gli Autori riscontrano un effetto positivo dell'aumento del salario minimo sul prezzo dei prodotti venduti nei fast food del New Jersey. Cioè, il costo dell'aumento del salario minimo è stato almeno in parte pagato dai consumatori.

Figura 8
Tassi di partecipazione a un tirocinio per i soggetti inclusi, rispettivamente, nel gruppo di trattamento e nel gruppo di controllo



Fonte: Martini *et al.*, 2018

rimentano alcuni tra i soggetti inclusi nel gruppo di controllo.

Il caso di *Lavoro&Psiche* esemplifica in modo nitido il classico problema che si pone frequentemente in un RCT: anche se – come in questo caso – l'assegnazione casuale fa la differenza per l'effettiva esposizione all'intervento – il tirocinio, in questo caso – difficilmente la determina in modo univoco.

Ne segue un dilemma:

- se si confrontano i due gruppi di soggetti con/senza *job coach,* si sfruttano appieno i benefici della randomizzazione, cioè si confrontano due gruppi di soggetti per costruzione in tutto e per tutto confrontabili. Ma questo confronto *non* identifica l'effetto della partecipazione a un tirocinio perché la disponibilità del *job coach* – assegnata casualmente – *non* determina univocamente l'esperienza di un tirocinio;
- d'altra parte, nemmeno il confronto dei due gruppi che hanno/non hanno sperimentato un tirocinio identifica l'effetto del tirocinio, perché i due gruppi *non* sono determinati casualmente. Vale a dire che, in generale, potrebbero differire anche rispetto ad altre caratteristiche oltre all'esperienza del tirocinio.

La via di uscita – anche in questo caso non priva di controindicazioni – è data dal metodo delle *variabili strumentali*:

- si parte dal confronto tra i due gruppi selezionati mediante randomizzazione. In questo modo si identifica l'effetto sugli esiti occupazionali di avere avuto a disposizione un *job coach*. In letteratura è chiamato effetto *Intention to Treat* (ITT);
- si riscala l'ITT per tenere conto del fatto che nel corso del periodo sperimentale tra coloro che hanno avuto a disposizione un *job coach* solo il 55% ha avuto accesso a un tirocinio, mentre tra coloro che non hanno avuto a disposizione un *job coach* il 25% ha comunque trovato il modo di accedere a un tirocinio:

impatto del tirocinio = impatto della disponibilità di un *job coach*/(0,55-0,25).

Applicata al caso considerato, si ottiene un *effetto medio del tirocinio* sulla *probabilità di lavorare almeno un giorno* nel corso del primo anno successivo alla sperimentazione pari a 19 punti percentuali. L'effetto medio sul *numero di giorni di lavoro* è pari a circa 50.

Questa soluzione sfrutta i benefici della randomizzazione – l'equivalenza tra i due gruppi selezionati casualmente – apportando *ex post* una correzione per tenere conto del fatto che l'effettiva partecipazione al tirocinio non è determinante univocamente dall'esito della randomizzazione. Il prezzo di questa soluzione in generale non è trascurabile. Imbens e Angrist (1994) hanno mostrato che in questo modo si ottiene un effetto medio su un particolare sottogruppo di soggetti, i cosiddetti *compliers*. Nel caso qui considerato si tratta dei soggetti che accedono al tirocinio *se e solo se* vengono assistiti da un *job coach*. Vale a dire che l'effetto medio trovato in questo modo in generale *non* può essere generalizzato a coloro che sono in grado di accedere al tirocinio anche in assenza di *job coach*, né a coloro che non accedono al tirocinio nemmeno se assistiti dal *job coach*[3].

Nel caso qui considerato, gli autori producono evidenze convincenti che il risultato ottenuto per i *compliers* può essere generalizzato anche agli altri soggetti (si veda l'articolo per i dettagli). Ma si tratta di un caso particolarmente favorevole. In generale, i *compliers* sono diversi dal resto della popolazione, per cui l'effetto medio che vale per loro non può essere generalizzato all'intera popolazione.

La lezione generale che si può trarre da questo caso è che, per quanto il valutatore faccia del suo meglio per tenere sotto controllo la composizione dei due gruppi di soggetti destinati all'esposizione/mancata esposizione all'intervento, gli altri attori coinvolti nell'implementazione dell'intervento – a partire dai soggetti inclusi nello studio – hanno almeno in parte il potere di modificare le scelte fatte dal valutatore.

4. Considerazioni conclusive

La logica controfattuale, qui introdotta per risolvere il problema dell'identificazione degli effetti di un intervento, è in realtà alla base del metodo scientifico di identificazione delle relazioni di causa ed effetto, quale che sia il particolare ambito disciplinare nel quale si colloca lo studio. Ad esempio, Diamond (1997) formula nitidamente nel linguaggio controfattuale la sua ipotesi fondamentale che a causare le diseguaglianze nel grado di sviluppo delle varie parti del globo siano le loro caratteristiche geografiche e non differenze biologiche tra le varie popolazioni:

> I expect that if the populations of Aboriginal Australia and Eurasia could have been interchanged during the Late Pleistocene, the original Aboriginal Australians would now be the ones occupying most of the Americas and Australia, as well as Eurasia, while the original Aboriginal Eurasians would be the ones now reduced to downtrodden population fragments in Australia.

È ancora Diamond (1997) a insistere che è questa la logica da adottare per risolvere un problema di identificazione di una relazione causale, si sia o meno nelle condizioni di svolgere un RCT:

> But laboratory experimentation can obviously play little or no role in many of the historical sciences. One cannot interrupt galaxy formation, start and stop hurricanes and ice ages, experimentally exterminate grizzly bears in a few national parks, or re-run the course of dinosaur evolution … Instead, one must gain knowledge in these historical sciences by other means, such as *observation*, *comparison*, and so-called *natural experiments* … How can students of human history profit from the experience of scientists in other historical sciences? A methodology that has proved useful involves the so-called natural experiments … While neither astronomers studying galaxy formation nor human historians can manipulate their systems in controlled laboratory experiments, they both can take advantage of natural experiments, by *comparing systems* differing in the *presence or absence of some putative causative factor*. [corsivi dell'Autore].

Se praticabile, RCT è senza dubbio il modo più affidabile per ottenere stime credibili dell'effetto causale di un intervento. Ma… serve essere preparati a procedere in modo osservazionale per due distinti motivi:

- primo, anche nel RCT meglio disegnato è facile che succeda qualcosa di non previsto dal valutatore – e fuori dal suo controllo – che in qualche misura al-

3 Imbens e Angrist (1994) mostrano che condizione necessaria per la validità di questo risultato è l'assenza di soggetti *defiers*. Nel caso considerato, sono soggetti che riescono ad accedere al tirocinio se non sono assistiti dal *job coach*, ma non vi riescono se assistiti. In molte circostanze, questa inclusa, si tratta di una condizione del tutto plausibile.

tera il protocollo sperimentale, rendendo necessarie correzioni *ex post*. Il caso discusso nel punto d) del paragrafo 3 esemplifica in modo efficace;

- in secondo luogo, ci sono miriadi di domande interessanti riguardanti gli interventi pubblici (e più in generale, le relazioni causali nelle scienze sociali) e molte meno opportunità di condurre RCT: sarebbe un peccato non dare risposta a quelle domande solo perché non c'è modo di condurre un RCT.

Ma anche quando non è possibile condurre un RCT, è importante avere chiara la distinzione tra valutazioni prospettiche e valutazioni retrospettive. Sono dette prospettiche le valutazioni progettate *congiuntamente* – o quanto meno in parallelo – all'intervento oggetto di valutazione. Cioè le valutazioni per le quali vengono predisposte per tempo le condizioni per il loro svolgimento. Sono invece dette retrospettive le valutazioni disegnate *dopo* che l'intervento oggetto di valutazione si è concluso.

Il terzo e il quarto caso presentati nel paragrafo precedente – uno con randomizzazione, l'altro osservazionale, si noti – sono esempi notevoli di valutazione prospettica. In entrambi i casi i due gruppi – esposti e non esposti – sono stati identificati dai valutatori *prima* dell'inizio dell'intervento. In entrambi i casi la regola di selezione è trasparente: assegnazione casuale in un caso, discontinuità geografica nell'altro. In entrambi i casi, sono state predisposte per tempo le condizioni per limitare il rischio della mancata comparabilità dei due gruppi: nel caso di *Lavoro&Psiche* (par. 3 punto d) selezionando casualmente i due gruppi; nel caso della riforma del salario minimo (par. 3 punto c) selezionando in modo ragionevole – prossimità geografica – il gruppo di confronto e rilevando *prima* dell'inizio dell'intervento le caratteristiche delle unità coinvolte utili a stabilire il grado di comparabilità dei due gruppi.

I casi discussi nei punti a) e b) del paragrafo 3 sono invece esempi di valutazioni retrospettive: a intervento completato, il valutatore ha scoperto che erano disponibili i dati utili alla valutazione, in entrambi i casi dati di origine amministrativa. I dati di origine amministrativa possono risultare – come nei due casi detti – di grande utilità per la valutazione di un intervento. La loro ovvia limitazione è data dal fatto che si tratta di dati generati per le finalità amministrative proprie dell'istituzione che li detiene, non per finalità scientifiche. Se va bene, risolvono il problema del valutatore, se va male sono inutili. Nel qual caso, la valutazione retrospettiva risulta impraticabile.

In definitiva, per ridurre il rischio di fallire la valutazione – cioè di non essere in grado di identificare un adeguato gruppo di confronto – è di fondamentale importanza iniziare a studiarne la fattibilità non appena si inizia a progettare l'intervento oggetto di valutazione.

Bibliografia

Angrist J.D., Pischke J-S. (2014), *Mastering 'Metrics: The Path from Cause to Effect*, Princeton, Princeton University Press

Angrist J.D, Rokkanen M. (2015), Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff, *Journal of the American Statistical Association*, 110, n.512, pp.1331-1344

Battistin E., Rettore E. (2008), Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs, *Journal of Econometrics*, 142, n.2, pp.715-730

Bloom H.S. (2006), *The Core Analytics of Randomized Experiments for Social Research, MDRC Working Papers on Research Methodology*, MDRC <https://bit.ly/2T6BihE>

Card D., Krueger A.B. (1994), Minimum Wages and Employment. A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84, n.4, pp.772-793

Card D., Krueger A.B. (2000), Minimum Wages and Employment. A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. Reply, *American Economic Review*, 90, n.5, pp.1397-1420

Diamond J. (1997), *Guns, germs and steel. The Fates of Human Societies*, New York, Norton&Co

Duflo E., Banerjee A. (2017), *Handbook of Field Experiments*, Amsterdam, Elsevier

Imbens G.W., Angrist J.D. (1994), Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, n.2, pp.467-475

Martini A., Rettore E., Barbetta G. (2018), *The Impact of Traineeships on the Employment of the Mentally Ill. The Role of Partial Compliance*, IZA Discussion Papers n.10582, Bonn, IZA

Pinotti P. (2017), Clicking on Heaven's Door. The effect of immigrant legalization on crime, *American Economic Review*, 107, n.1, pp.138-168

Enrico Rettore

enrico.rettore@unitn.it

È Professore di Econometria presso l'Università degli Studi di Trento, in precedenza Professore di Statistica economica presso l'Università degli Studi di Padova, dove ha ottenuto il Dottorato in Statistica. Il suo principale interesse di ricerca è la valutazione degli effetti di politiche pubbliche nel più ampio contesto dell'inferenza causale basata su campioni auto-selezionati. Su questi temi ha pubblicato vari articoli in riviste del settore quali ad esempio la *Review of Economics and Statistics*, l'*American Economic Review*, il *Journal of Econometrics* e il *Journal of the Royal Statistical Society* (A). Ha diretto vari progetti di ricerca finanziati da ministeri italiani e dalla Commissione europea.