ESERCIZI DI VALUTAZIONE: UN'ANALISI D'IMPATTO DEI CRITERI

Paolo Rossi – Dipartimento di Fisica dell'Università di Pisa

Dobbiamo necessariamente partire da una premessa, che potrebbe sembrare banale se non ci scontrassimo quotidianamente con l'evidenza del fatto che purtroppo quasi mai ne vengono tratte le debite conseguenze: i criteri di valutazione della ricerca non dovrebbero in alcun modo prescindere, caso per caso, dalla definizione dei soggetti e degli oggetti della valutazione, dalle sue finalità e perfino dalla metodologia valutativa adottata.

L'adozione di criteri sostanzialmente uniformi per la valutazione di individui o di grandi aggregazioni organizzative, per obiettivi di verifica o di premialità, per misure di tipo quantitativo o qualitativo non può che produrre descrizioni inadeguate e distorte di almeno alcuni, se non di tutti, i contesti di applicazione dei criteri stessi.

In particolare occorrerebbe sempre distinguere radicalmente le valutazioni finalizzate a stabilire un qualche tipo di graduatoria da quelle volte a verificare il superamento di prefissati obiettivi minimi di produttività scientifica, quale che sia, in entrambi i casi la finalità dell'azione valutativa.

Ma soprattutto ci appare indispensabile che sia sempre mantenuta una profonda distinzione tra le azioni volte a esprimere un qualche tipo di giudizio sui risultati conseguiti da soggetti singoli da quelle mirate a rilevare l'efficienza e l'efficacia dei comportamenti collettivi di un'istituzione, sia essa un Dipartimento, un Ateneo o un Ente di ricerca, e che questa distinzione si rifletta in modo netto nella selezione dei criteri adottati, che non possono in alcun modo essere riportati a un comune denominatore.

Notiamo infatti che, mentre nelle valutazioni aggregate l'adozione di metodologie a base sostanzialmente statistica può, seppure soltanto con adeguate e non banali calibrazioni, produrre risultati significativi in virtù della legge dei grandi numeri, viceversa l'impiego acritico di parametri statistici nelle valutazioni individuali è di per sé un fondamentale errore metodologico, le cui conseguenze, che nel seguito cercheremo di esaminare in dettaglio, risultano prevalentemente e inevitabilmente negative.

Nell'analisi dell'impatto di metodi di valutazione di tipo statistico occorre necessariamente partire da una disamina dei principali errori metodologici, che possono rendere inattendibile il risultato anche quando esso sia il frutto dell'esame di un grande numero di dati, la cui stessa quantità parrebbe dover assicurare (ma purtroppo non assicura affatto) il valore conoscitivo dell'operazione valutativa effettuata.

Il primo e più grave errore che si potrebbe commettere (e che purtroppo viene commesso) in un'analisi di tipo statistico consiste in ciò che nel linguaggio tecnico va sotto il nome di *sampling bias*. Questo errore viene commesso ogni qual volta un campione è raccolto in modo tale da far sì che alcuni elementi della popolazione da campionare abbiano minor probabilità di altri di essere inclusi. Un esempio banale è quello delle interviste stradali, che per definizione escludono chiunque non abbia la possibilità o la voglia di andarsene in giro, e quelli che si spostano soltanto in auto: niente di male se l'obiettivo è conoscere l'opinione dei cittadini sullo stato dei marciapiedi, ma in quasi tutti gli altri casi il campione ha un *bias.*

Nel caso che ci interessa, vale la pena di confrontare le procedure adottate per la VTR (2001-2003) con quelle della VQR (2004-2010 e 2011-2014). Nel primo caso si valutava un numero di oggetti individuato dalle sedi e proporzionale al numero dei ricercatori, mentre nel secondo caso si valutava un numero di oggetti prefissato per ciascun ricercatore, ma ricordiamo che in entrambi i casi l'obiettivo dichiarato era quello di valutare la qualità *complessiva* dell'istituzione (nello specifico del Dipartimento). Ma è evidente

che nella VTR l'intento era quello di restringere l'attenzione alla produzione scientifica di qualità medio-alta, e quindi il *bias*, che pure esisteva (in quanto i lavori di scarsa qualità non venivano di regola proposti per la valutazione), era dichiarato e in qualche modo condiviso anche dai valutati, perfettamente consapevoli che il giudizio su una parte rilevante (quella più seriale) della produzione non poteva raggiungere significativi livelli di qualità, nel secondo caso la valutazione, che a parole si voleva collettiva, era di fatto individuale, in quanto si escludevano dalla valutazione prodotti anche ottimi di ricercatori qualificati mentre erano di necessità inclusi per "far numero" i lavori minori di ricercatori demotivati. Anche se formalmente il criterio della proporzionalità tra le dimensioni dell'istituzione e il numero di prodotti richiesti era rispettato in entrambi i casi, nel caso della VQR era tecnicamente impossibile estrarre dai risultati un giudizio sul valore *assoluto* del Dipartimento, emergendone al più tanto una fotografia della distribuzione qualitativa (o spesso soltanto anagrafica) dei suoi componenti.

Non a caso quindi, mentre il risultato della VTR nella maggior parte dei casi finiva con il collimare con la percezione del valore relativo e assoluto delle diverse sedi che già possedevano i membri più avvertiti ed esperti delle varie comunità, viceversa gli esiti della VQR sono apparsi spesso paradossali, essendo facile per un'istituzione "giovane", a causa della metodologia di campionamento adottata, ottenere, con una produzione tutto sommato mediocre ma diffusa, risultati migliori di un'istituzione "vecchia" in cui i risultati "eccellenti" di alcuni potevano essere facilmente oscurati dalla scarsa e cattiva produzione di molti soggetti anziani e ormai largamente inattivi.

Un altro grave errore commesso nelle valutazioni di tipo statistico consiste nell'attribuire un peso importante, e spesso eccessivo, ad alcuni indicatori (spesso di natura bibliometrica) che fanno riferimento alla sede di pubblicazione dei risultati o al loro impatto citazionale.

Nel primo caso, sia che il criterio sia basato sull'*Impact Factor,* come avviene negli ambiti in cui le riviste sono indicizzate, sia che si fondi su una classificazione delle riviste stesse (Classe A), la sua applicazione si fonda sull'assunto, fondamentalmente errato, che la qualità, o comunque l'impatto citazionale, degli articoli pubblicati su una rivista scientifica si distribuisca intorno a un valore medio e con una limitata varianza. In realtà molti autorevoli studi hanno dimostrato che la distribuzione delle citazioni degli articoli pubblicati su una stessa rivista ubbidisce a una legge di potenza, tipica delle distribuzioni prive di scala, per la quale la media ha scarso significato statistico e la varianza non ne ha alcuno. Per dirla in termini rozzi ma efficaci, la maggior parte degli articoli pubblicati su una rivista ad elevato *Impact Factor* vive di luce riflessa, avendo assai limitata visibilità (e interesse) individuale ma beneficiando della grande visibilità e interesse di alcuni articoli "civetta", il cui fortissimo impatto individuale viene "spalmato" su tutta la rivista. Sarebbe come dire, facendo la media delle battaglie vinte, che ogni soldato di Napoleone sarebbe stato un buon generale.

A questo paradosso si sono immaginati alcuni correttivi, come il tener conto della posizione relativa, in termini citazionali, del singolo articolo rispetto alla media delle citazioni della rivista in cui è pubblicato. Ma anche questo approccio rivela la sua parziale fallacia, in quanto le riviste scientifiche coprono quasi sempre un numero non esiguo di differenti ambiti e sottoambiti disciplinari, e il numero delle citazioni non ha un significato assoluto, ma andrebbe sempre commisurato alle dimensioni della comunità (o sottocomunità) di riferimento. Sembra evidente che il numero delle citazioni che si possono conseguire in una piccola comunità disciplinare è intrinsecamente limitato, quale che sia la qualità dei risultati presentati, per cui a meno che non si teorizzi che le piccole dimensioni di una specifica comunità disciplinare siano in sé un difetto da penalizzare occorrerebbe, in qualsiasi analisi citazionale, tener conto non tanto del numero medio di citazioni della sede editoriale quanto del numero medio di citazioni nell'ambito del sottogruppo

disciplinare, operazione peraltro estremamente difficile, anche se vi sono indicazioni di comportamenti "universali" (in senso tecnico) della distribuzione delle citazioni una volta normalizzata al numero medio delle citazioni (e delle pubblicazioni) nello specifico ambito di riferimento.

Abbiamo già accennato all'effetto deleterio derivante dall'attribuzione di un qualunque peso (incluso lo zero) alla presenza nell'istituzione di soggetti inattivi, ma ci preme sottolineare che tale effetto risulta particolarmente aggravato dall'utilizzo dei risultati della valutazione nell'attribuzione di risorse, umane e finanziarie, in quanto la mancata o minore attribuzione non penalizza minimamente i soggetti inattivi, che per definizione non hanno bisogno di risorse, e nei confronti dei quali le istituzioni non hanno a disposizione alcuno strumento atto a sollecitarne una maggior produttività, mentre ovviamente ne risultano danneggiati i soggetti attivi, la cui produttività potrebbe ridursi proprio in conseguenza della scarsità di risorse, con un *feedback* negativo che nel tempo porterebbe soltanto a un ulteriore peggioramento delle situazioni in cui sono già presenti criticità.

Negli esercizi della VQR ci sarebbe da segnalare anche l'emergere di diversi altri problemi, più o meno gravi, di natura metodologica, quali l'erronea pratica di sommare percentili o una scarsa attenzione alla dipendenza quantitativa della varianza dalla dimensione dei campioni, per cui ad esempio è abbastanza inevitabile che una grandissima istituzione, come la Sapienza di Roma, finisca, a causa del suo stesso peso statistico (7% dell'intero sistema universitario) per appiattirsi su valori medi che essa stessa contribuisce pesantemente a determinare. Appare abbastanza grave anche la mancanza di un criterio di normalizzazione atto a calibrare adeguatamente il peso dei lavori scritti in collaborazione, con il rischio di una loro sopravvalutazione, quando fossero contati tante volte quanti gli autori, di una sottovalutazione, quando fossero contati una sola volta, o addirittura di una non valutazione. Non torneremo comunque in questa sede su queste e su altre possibili critiche, che sono già state abbondantemente discusse altrove: dispiace soltanto che non siano state finora affrontate dai valutatori con tutta l'attenzione che meriterebbero, mentre sono state talvolta adottate misure palliative che non toccano l'essenza dei problemi e in certi casi giungono addirittura ad aggravarli.

Esiste invece a nostro parere un correttivo al *sampling bias* che, se accompagnato da alcune cautele metodologiche, potrebbe rappresentare una via d'uscita, almeno temporanea, da alcuni dei paradossi sopra evidenziati, e che malgrado la sua apparente radicalità sarebbe anche perfettamente in linea con una legislazione già esistente. Si tratterebbe infatti di dare concreta applicazione all'Art. 3-bis della Legge 9 gennaio 2009, n.1, di conversione con modificazioni del D.L. 10 novembre 2008, n.180, che prevede l'istituzione dell'ANPRePS (Anagrafe Nazionale dei Professori e dei Ricercatori e delle Pubblicazioni Scientifiche), con la creazione di un *database* proprietario e certificato dell'intera produzione scientifica nazionale, che permetterebbe di effettuare la valutazione della ricerca su tutta la produzione, eliminando tutti i *bias* da campionamento e sfruttando la legge dei grandi numeri per attenuare, almeno statisticamente, gli effetti dovuti alle sopra elencate difficoltà di attribuzione di un valore qualitativo che sia al tempo stesso assoluto ed oggettivo . Nulla impedirebbe un cauto dosaggio di indicatori di tipo bibliometrico, ma l'ipotesi di base resterebbe quella che, nel complesso della produzione (molte centinaia di migliaia di prodotti), la frequenza relativa dei prodotti buoni, mediocri e scarsi sia abbastanza uniforme tra le varie sedi.

Merita a tale proposito sottolineare che un recentissimo documento (*Aggiornamento 2017 al Piano Nazionale Anticorruzione*) diffuso dall'ANAC dedica un'ampia disamina all'importanza e urgenza di giungere all'effettiva attuazione dell'ANPRePS, anche con precise finalità di trasparenza e imparzialità delle valutazioni. Difficile comprendere che cosa osti a dar seguito a questa indicazione.

A parziale conclusione di questa parte dell'analisi vogliamo mettere in evidenza quelli che a nostro avviso sono i principali effetti negativi di una valutazione delle istituzioni condotta nelle forme e nei modi che conosciamo attualmente nel nostro Paese.

Da un lato si assiste alla creazione di graduatorie largamente artificiali (*ranking*) che spesso esaltano differenze minori, talvolta oscurando differenze sostanziali. Non è detto che tra il primo e il quinto di una graduatoria di dieci ci sia una reale e oggettivamente distinguibile distanza, mentre può benissimo accadere che intercorra un vero distacco tra il quinto e il sesto di quella stessa graduatoria, ma l'opinione pubblica percepisce soltanto la posizione in graduatoria e non è messa in grado di percepire prossimità e distanze reali. Già una presentazione basata su pochi *cluster* abbastanza omogenei (*rating*) rappresenterebbe un progresso sostanziale su questo fronte.

D'altro canto, e questo ci pare l'effetto più grave, questo tipo di valutazione, abbinato a logiche esaltate come premiali ma in realtà soltanto penalizzanti nella ripartizione delle risorse, finisce col danneggiare ulteriormente le realtà più deboli (*Matthew effect*), togliendo loro di fatto ogni possibilità di recupero.

Passando all'ancor più impegnativo tema dei criteri per la valutazione individuale, che incidono sia sui processi abilitativi, sia sui concorsi veri e propri, notiamo innanzitutto che i criteri puramente statistici sono in questo caso *sempre* inadeguati, in quanto , come già argomentato in precedenza a proposito degli indicatori bibliometrici, si basano sulla media di situazioni spesso non confrontabili, anche ammesso e non concesso che un confronto con valori medi possa contribuire significativamente alla formulazione di un giudizio individuale, e comunque non sono assolutamente in grado di cogliere alcun elemento relativo alla reale qualità dei risultati presentati.

A questa grave limitazione si deve aggiungere l'errore di stabilire "soglie" di produttività scientifica determinate meccanicamente sulla base di medie o mediane, o anche soltanto di percentili, spesso aggregando i dati di sottocomunità scientifiche tra loro non omogenee, e senza riferirsi agli *standard* che tali comunità e sottocomunità già condividono al loro interno, né tantomeno tener conto del fatto che la diversa storia e reputazione di un ambito disciplinare può far sì che un valore definito a partire da una media possa risultare penalizzante per una comunità dall'elevato profilo scientifico, anche internazionale, e invece indebitamente premiante per una nicchia accademica nel suo complesso scarsamente produttiva.

Una valutazione concettualmente inadeguata a livello individuale può produrre (e in taluni casi ha già prodotto) effetti realmente devastanti. Ne elenchiamo soltanto alcuni. Il primo e più grave rischio consiste nel privilegiare la ricerca *mainstream* (più facile e più riconosciuta) a scapito delle linee di ricerca meno frequentate e soprattutto della ricerca veramente originale, che (la storia insegna) quasi sempre fatica parecchio a farsi riconoscere in un contesto di "scienza normale" (nel senso di Kuhn). Un'altra tipica vittima di questo meccanismo è la multidisciplinarità, che soprattutto in un mondo accademico attestato (per colpa non soltanto sua, ma anche della scarsità di risorse) su una difesa arroccata delle proprie sfere d'influenza e poco disposto a lasciar spazio ad aperture verso altre realtà.

Procedure di valutazione largamente meccanica tendono poi a stimolare vari tipi di comportamenti adattivi ed elusivi, che non si limitano soltanto a forme di produzione seriale, ma spaziano dalle false collaborazioni ai *network* citazionali, da strumentali cambi di settore disciplinare a procedure di "chiamata diretta" spesso giocate su aspetti formali (equipollenze non sempre evidenti con posizioni accademiche estere) piuttosto che su una reale verifica di competenze, perpetuando con muove modalità il tanto spesso vituperato "nepotismo accademico".

E ancora: un malinteso concetto di originalità della produzione scientifica (che peraltro non impedisce in alcun modo la produzione seriale) ha portato a escludere dalla valutazione i prodotti di alta divulgazione scientifica, a cominciare dai manuali, che sono stati e sono ancora spesso uno dei principali strumenti di trasmissione delle conoscenze da generazione a generazione, ma la cui produzione è oggi largamente disincentivata, e non solo tra i ricercatori più giovani, a causa dell'assenza di peso di questo tipo di produzione (non certo facile né banale) nel quadro della valutazione della ricerca scientifica.

E da ultimo (*last but not least*): non si manifesta alcuna attenzione per le problematiche di genere nella produzione scientifica. Numerosi, e ormai ben studiati, sono i meccanismi che costituiscono uno svantaggio competitivo per le donne nel mondo della ricerca, in particolare in alcune discipline, e ben noto è il meccanismo del "soffitto di cristallo" che le vede pesantemente penalizzate nei processi di progressione di carriera. Ma non si vede come gli automatismi della valutazione basata su indicatori statistici possano essere soggetti a correttivi efficaci al fine di contrastare questo tipo di inaccettabile discriminazione.

A fronte di tutte queste difficoltà ci sembra che l'unico possibile e serio correttivo consista in una sostanziale rivalutazione del "giudizio dei pari", che può benissimo tener conto degli esiti degli indicatori quantitativi (*informed peer review*), ma deve comunque farlo alla luce della competenza scientifica e del giudizio di merito qualitativo che soltanto un esperto valutatore può formulare.

A tale proposito occorrerebbe anche ripensare i meccanismi di formazione delle commissioni giudicatrici, che da un lato non possono essere affidate soltanto al capriccio della sorte, estromettendo le comunità scientifiche da un processo di autoselezione capace di individuare al proprio interno i soggetti più capaci e più adatti per questo delicato compito. All'inevitabile rischio di comportamenti non conformi al mandato si potrebbe in qualche misura ovviare con meccanismi sanzionatori (quali l'esclusione da future commissioni), basati sulla verifica *ex post*, dopo un congruo lasso di tempo, della qualità delle selezioni effettuate.

Dall'altra parte sarebbe importante anche ripensare seriamente i meccanismi di esclusione *ex ante* dalle funzioni di commissario, che oggi sono legati soprattutto a una verifica della produzione scientifica *recente*, giungendo al paradosso, anche umanamente umiliante, di escludere dalle commissioni soggetti di elevatissima competenza scientifica che per i più svariati e comprensibili motivi (impegno didattico e gestionale *in* primis) possono aver ridotto la quantità della propria produzione, per di più non necessariamente a scapito della qualità.

Infine bisognerà anche cominciare a pensare a un futuro della valutazione nel quale i criteri attualmente in uso possano essere abbandonati in favore di strumenti più moderni e sensibili per l'analisi dell'impatto dei risultati della ricerca scientifica. Riteniamo che i continui progressi della *data science*, uniti all'evidenza che la produzione scientifica nel suo complesso ha a tutti gli effetti la struttura dei *big data*, potranno consentire in un tempo ormai prossimo, in particolare grazie alla cosiddetta *sentiment analysis*, una misura dell'effettiva risposta delle comunità scientifiche ai nuovi risultati e alle nuove idee che prescinda da strumenti e indicatori così artificiali e purtroppo anche così manipolabili come quelli oggi utilizzati.