Dibattito in merito alle procedure e agli esiti della Valutazione della Qualità della Ricerca (VQR 2004-2010)

Interventi di:
Gaetano Domenici
Marco Catarci
Rosa Capobianco
Valeria Biasi

Valutazione della Qualità della Ricerca

Affidabilità dei dati, delle procedure, dei giudizi valutativi e degli esiti conoscitivi; equità, merito e distribuzione delle risorse: perché se ne discute (invano, o quasi) dal 1998?

Gaetano Domenici

Università degli Studi «Roma Tre», Dipartimento di Scienze della Formazione

gaetano.domenici@uniroma3.it

L'Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) ha presentato, lo scorso mese di luglio, i risultati della procedura di Valutazione della Qualità della Ricerca italiana per il periodo 2004-2010.

È la seconda volta che in Italia è stato compiuto un esercizio valutativo della ricerca svolta in strutture pubbliche (*in primis* Università e enti come il CNR). Entrambe le volte, soprattutto dopo la presentazione degli esiti, il dibattito, non solo tra gli addetti al lavoro, si è acceso con grande vivacità.

Inizialmente ha avuto come fulcro la positività o meno di un processo valutativo in un settore, come quello della ricerca, ritenuto per molti versi imponderabile perché assai connesso alla creatività individuale o di gruppi di ricercatori (per non pochi anni, come si sa, la valutazione da molti auspicata è stata da quasi tutti temuta); poi, il dibattito si è spostato sull'efficacia reale di provvedimenti premiali, sul piano finanziario, di strutture e/o soggetti ritenuti migliori sulla base di procedure non accettate da tutti, perché , secondo molti, per una auspicabile giustizia distributiva, si dovrebbero finanziare, soprattutto quelle strutture di ricerca che, o per dislocazione geo-economica o per aree e settori scientifico-disciplinari di riferimento, non possono accedere a più e più cospicue fonti finanziarie. Salvo poi, in questi casi, penalizzarle in presenza di un non miglioramento degli esiti della ricerca nonostante il credito concesso. Continuando, in una pur breve carrellata di questioni di non poco conto, nella discussione ci si è successivamente soffermati sull'affi-

dabilità degli esiti di un processo valutativo fondato su parametri, procedure e criteri valutativi complessivamente non condivisi dall'intera comunità scientifica, ovvero sulla malcelata soggettività dei giudizi. Infine, gli elementi più raffinati del dibattito politico e culturale sembrano essere quelli che in termini strategici, andando oltre le mere contingenze, puntano all'analisi critica del cambiamento sia di atteggiamento, sia strutturale della ricerca che si sta producendo, soprattutto in ambito universitario, anche attraverso la legislazione valutativa, che quindi si soffermano nella pre-figurazione del quadro complessivo che si sta producendo attraverso le decisioni di politica della ricerca di ieri e di oggi.

In effetti, come mi è capitato di scrivere su questo Journal, i più recenti interventi di legge relativi all'Università italiana in particolare riferiti *alla valutazione della ricerca*; alla *organizzazione dell'Università, con riferimento al personale accademico e al reclutamento, nonché alle norme per incentivare la qualità e l'efficienza del sistema universitario*, stanno dando origine ad una riorganizzazione strutturale della ricerca in ambito universitario.

Le norme e le proposte normative, compreso il nuovo sistema di reclutamento dei docenti, in qualche modo postulano e comportano, infatti, un'accentuata modificazione delle forme organizzativo-procedurali della ricerca universitaria (oltre che delle strutture pubbliche di ricerca), nonché degli strumenti e delle finalità della sua valutazione.

La prevista *riduzione a circa la metà degli attuali 370 settori scientifico-disciplinari, (con una consistenza minima di 50 ordinari per settore); la soppressione delle Facoltà*, con il superamento della separazione delle strutture deputate allo svolgimento della didattica e della ricerca; l'irrobustimento delle procedure non solo di autovalutazione, ma anche e soprattutto di valutazione esterna della produttività scientifica tanto dei docenti universitari (in particolare della loro attività di ricerca, con la perdita, in caso di valutazione negativa, dello scatto stipendiale e del diritto di partecipazione alle commissioni concorsuali), quanto delle aree disciplinari in cui viene formalmente articolato il sapere, e delle strutture universitarie, con la distribuzione di parte delle risorse in base alla qualità della ricerca; le procedure concorsuali con nuovi criteri valutativi della produzione scientifica dei candidati, eccetera, sono, a ben vedere, taluni degli elementi di novità che stanno producendo effetti di enorme portata sia sui processi di organizzazione e messa in atto della ricerca, sia su quelli valutativi della stessa, così orientando non poco atteggiamenti, disposizioni nonché ambiti, tipologia e finanziamento delle stesse attività di ricerca (soprattutto nell'area 11 CUN che interessa non poco, anche come «oggetto» di ricerca, questo Journal).

Dati di fatto, questi ultimi, cui si associano da una parte, il basso numero di ricercatori per numero di abitanti, e un investimento in percentuale

del PIL tra i più bassi della UE; dall'altra, il pericolo gravissimo di creare condizioni di soffocamento, se non morte della ricerca in settori scientifico-disciplinari numericamente minoritari, a causa dei nuovi accorpamenti CUN e dipartimentali, e dei finanziamenti dei progetti di ricerca che sempre più privilegiano, fino ad esclusivizzarlo, il momento ex-ante della valutazione, cioè quello che può orientare più politicamente che scientificamente le scelte di gruppi e strutture di ricerca. E ciò, soprattutto, in presenza della drastica riduzione dei finanziamenti pubblici e delle risorse umane, in un Paese, come il nostro nel quale anche gli investimenti finanziari nella ricerca da parte del mondo industriale risultano tra i più bassi dell'area OCSE.

Un rischio, questo, ancor più esiziale se si considera, in estrema sintesi, che ad una simile perdita di indipendenza dell'Università è assai probabile che corrisponda una ancor più grave perdita del rilievo della Ricerca generale, disinteressata, la più produttiva a medio e lungo termine, a vantaggio di quella troppo finalizzata e applicativa che, pur importante, rappresenta una risposta alle sollecitazioni – soprattutto finanziarie – di un mondo, anche in campo valutativo (si pensi alla estensione spesso arbitraria dell'impiego dell'*Impact Factor* in ambiti non pertinenti) che non sempre ha a cuore le «magnifiche sorti e progressive della scienza».

Si consideri che l'avvio della valutazione della ricerca in Italia è rintracciabile, sul piano ordinamentale, nel Decreto Legislativo 204 del 5 giugno 1998, che aveva dotanto il nostro paese, seppur tardivamente, di uno strumento assai utile per il miglioramento del sistema nazionale della ricerca. Gli articoli 4 e 5 di quel decreto, prevedevano infatti, rispettivamente:

a. L'istituzione di organi rappresentativi della comunità scientifica nazionale (CSN: Consigli Scientifici Nazionali), i quali in collaborazione con i rappresentanti della società civile (amministrazioni pubbliche, mondo della produzione, dei servizi, eccetera) dovevano costituire l'Assemblea della Scienza e della Tecnologia (AST) con lo scopo, tra l'altro, di formulare «osservazioni e proposte per l'elaborazione e l'aggiornamento del PNR» (Programma Nazionale per la Ricerca), per contribuire a determinare (art. 1) gli «indirizzi e le priorità strategiche per gli interventi (del Governo) a favore della ricerca scientifica e tecnologica, definendo il quadro delle risorse finanziarie».
b. La costituzione del Comitato di Indirizzo per la Valutazione della Ricerca (CIVR) con lo scopo di operare «per il sostegno alla qualità e alla migliore utilizzazione della ricerca scientifica e tecnologica nazionale, secondo autonome determinazioni con il compito di indicare i criteri generali per le attività di valutazione dei risultati della ricerca, di promuovere la sperimentazione, l'applicazione e la diffusione di metodologie, tecniche e pratiche di valutazione […], favorendo al riguardo il confronto e la coo-

perazione tra le diverse istituzioni operanti nel settore, nazionali e internazionali».

Dell'art. 5 del Decreto, in particolare, era stato apprezzato il collegamento esplicito del processo valutativo con la promozione del *confronto* e – soprattutto – della *cooperazione* tra le istituzioni di ricerca che operano nello stesso settore o nella medesima area; confronto e cooperazione, indispensabili per la costituzione degli organi rappresentativi della comunità scientifica nazionale, compresa ovviamente quella universitaria, cui faceva riferimento il citato art. 4, e utili a favorire un più forte sviluppo e una più ampia *diffusione della cultura della valutazione.* È accaduto invece, che l'attuazione di quanto previsto dall'art. 5 (costituzione del CIVR e attuazione del processo di valutazione della ricerca) è avvenuta senza dar corso a quanto contemplato dall'art. 4 (costituzione dei Consigli Scientifici Nazionali e Assemblea della Scienza e della Tecnologia), con i conseguenti problemi di scarsa o nulla condivisione delle modalità di definizione e applicazione dei criteri valutativi dei prodotti della ricerca relativi a tante aree e/o sub-aree disciplinari, che a dir poco indeboliscono la credibilità e gli effetti della valutazione condotta.

L'apporto che in tal senso avrebbero potuto dare i previsti Consigli Scientifici Nazionali, sarebbe stato di grande rilievo: come si sa, gli esperti di un dato settore, pur abbastanza ampio, possono più e meglio di altri contribuire alla definizione di criteri valutativi generali condivisibili (e alla determinazione della loro significatività) nelle ricerche interne o molto prossime a quel settore specifico.

In effetti, tutta la decretazione successiva e poi le procedure applicate, hanno purtroppo stravolto lo spirito e le finalità, e perciò l'alta coerenza interna del DLgs 204. Delle attività di «cooperazione e confronto» tra le strutture di ricerca, si è prescelta, in modo quasi esclusivo, quella del «confronto» tra esse. Anziché costruire un sistema valutativo caratterizzato dall'uso di criteri e procedure condivisi, in grado perciò di configurarsi come supporto a quelle strutture e come «sostegno alla qualità e alla migliore utilizzazione della ricerca scientifica», se ne è creato uno volto solo alla determinazione di una graduatoria (*ranking list*) delle strutture di ricerca e degli atenei, una vera e propria classificazione gerarchica del loro «presunto» valore, in rapporto al quale si dovrebbero correlare alcuni finanziamenti pubblici. Il tutto, peraltro, sulla base di una soggettività dei giudizi valutativi contrabbandati come oggettivi, perché non di rado espressi, nel primo esercizio di valutazione (CIVR), da *giudici inesperti o quasi-esperti* perché appartenenti ad aree conoscitive prossime ma non coincidenti con quelle relative ai prodotti di ricerca da valutare dei settori o sub-settori disciplinari dei prodotti della ricerca; e nel secondo esercizio di valutazione (VQR) da giudici non opportunamente preparati ad utilizzare scale di punteggi omogenee. L'effetto di compromissione della fiducia nella

valutazione giusta ed equa della ricerca, prodotto da queste modalità procedurali, ha contribuito e contribuisce non poco all'inverosimile rallentamento dello sviluppo della cultura della valutazione anche in ambito universitario.

Alla luce delle considerazioni appena fatte, si può ben comprendere perché oggi siano molto avvertiti nel mondo accademico alcuni problemi che solo parzialmente sembrano risolti da quanto previsto dal citato DM marzo/2010, Linee guida VQR 2004-2008, poi esteso al 2010, e dalla messa in pratica della VQR 2004-2010 i cui esiti sono stati presentati, come dicevo in apertura, nel luglio scorso, rispetto al primo esercizio di valutazione della ricerca 2001-2003. Sono, quelle appena fatte, considerazioni che si dovranno ri-approfondire – anche su questa rivista –, sia per meglio orientare le future scelte metodologiche, sia per mettere strutture e ricercatori in grado di rappresentare correttamente i prodotti delle proprie ricerche; sia ancora, per sviluppare strategie valutative teoricamente più consistenti e proceduralmente più affidabili di quelle a tutt'oggi disponibili e/o applicate.

C'è una questione certo non decisiva, ma neppure banale, del cambiamento dell'estensione dell'arco di tempo preso in esame dal secondo Esercizio, passato dai tre ai cinque anni, e poi portato a sette, cambiamento che non faciliterà i confronti diacronici delle strutture di ricerca e del Paese su basi di dati omogenee. Ma vi sono anche e soprattutto, questioni di fondo, relative, per esempio: ai criteri, sia quantitativi, sia qualitativi prescelti e definiti (dopo, non prima l'avvio dell'esercizio valutativo) compresa *la scala ad intervalli non regolari* per la collocazione valutativa dei «prodotti»; agli elementi ritenuti costitutivamente peculiari della qualità del prodotto-ricerca esaminato e della struttura di ricerca considerata; alla mancata considerazione ponderale della produttività complessiva di ricercatori e strutture; ai margini di errore valutativo ritenuto accettabile (le classifiche cambiano la posizione delle strutture se si accetta un benché minimo errore soggettivo di *un solo valutatore* nella strutturazione del giudizio *su una sola unità di prodotto*); infine, ma non ultimo, relative all'indipendenza di giudizio dei GEV di nomina ministeriale in un'Agenzia ritenuta indipendente, all'«addestramento» (mancato) per una corretta «normalizzazione» preventiva dei punteggi/giudizi assegnati dai valutatori. Una certa «opacità» dei dati non consente, peraltro, una affidabile più ricca ri-costruzione dell'itinerario percorso nell'esercizio valutativo.

Sulla base di queste e di altre considerazioni ancora, l'*ECPS Journal* ha preso le mosse per avviare e sostenere un più articolato dibattito scientifico e culturale sulla valutazione della ricerca nel nostro Paese.

In questo numero sono ospitati i primi tre interventi, rispettivamente di un rappresentante CUN per l'area 11, docente dell'area 11a; di un docente di statistica; e di un docente dell'area 11b, membro del Nucleo di Valutazione dell'Ateneo «Roma Tre».

L'esercizio della VQR 2004-2010 in area 11 e in ambito pedagogico

Marco Catarci

Università degli Studi «Roma Tre», Dipartimento di Scienze della Formazione

marco.catarci@uniroma3.it

1. Aspetti principali della procedura VQR

L'Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) ha presentato in data 16 luglio 2013 gli esiti della procedura di Valutazione della Qualità della Ricerca italiana (VQR) per il settennio 2004-2010 (ANVUR, 2013a). Pur nella consapevolezza della necessità di un'ampia e approfondita riflessione sui risultati dell'esercizio, sulle metodologie adottate e sulle implicazioni per il governo del sistema universitario, nel presente contributo ci si propone di disegnare sinteticamente un breve profilo della procedura, per poi evidenziare alcuni esiti relativi all'area 11, con particolare riferimento all'ambito pedagogico.

Il procedimento ha sottoposto a valutazione, tra il novembre 2011 e il giugno 2013, la qualità della ricerca di 133 strutture (95 università, 12 enti di ricerca vigilati dal Ministero per l'Istruzione, l'Università e la Ricerca e 26 enti di ricerca o consorzi interuniversitari «volontari») all'interno di 14 aree scientifiche (tra le quali la composita area 11, nella quale sono inclusi 38 settori scientifico-disciplinari appartenenti a diversi macro-settori scientifici: Antropologia, Filosofia, Geografia, Pedagogia, Psicologia, Storia).

Nel corso della procedura, 184.878 «prodotti» [1] di ricerca (articoli, monografie e saggi, atti di convegni, brevetti, manufatti, note a sentenza, traduzioni, software, banche dati, mostre e performance e cartografie) sono stati valutati in base ai seguenti criteri: rilevanza, originalità e grado d'internazionalizzazione.

[1] Il termine «prodotto» della ricerca viene qui utilizzato in quanto di uso comune in tema di valutazione della ricerca scientifica. Si ritiene opportuno, tuttavia, segnalare l'urgenza di una riflessione sull'adozione di un tale termine che, soprattutto nel caso della ricerca di base (i cui esiti non immediatamente applicabili rappresentano un contributo imprescindibile per la crescita culturale e l'innovazione del Paese), non appare del tutto appropriato.

I 14 Gruppi di Esperti della Valutazione (GEV) (uno per ciascuna area) hanno definito le modalità del procedimento, in particolare in riferimento all'adozione di indici bibliometrici (che tengono conto del numero di citazioni ottenute dagli articoli e dell'*Impact Factor* delle riviste) o di procedure di *peer-review* (attraverso l'analisi della qualità del singolo prodotto da parte di revisori anonimi, ma non in «doppio cieco»).

Per ogni struttura di ricerca, sono stati calcolati, infine, sette indicatori di area che esprimono la qualità dei prodotti di ricerca e dei processi di reclutamento, la capacità di attrarre risorse esterne e di creare collegamenti internazionali, la propensione alla formazione per la ricerca e all'utilizzo di fondi propri per finanziare la ricerca e il miglioramento della qualità scientifica rispetto all'esercizio di valutazione precedente. Sono stati altresì definiti otto indicatori legati alla cosiddetta «terza missione», che esprimono il grado di apertura della struttura di ricerca al contesto socio-economico, con attività di valorizzazione e trasferimento delle conoscenze (ad esempio, le attività di consulenza conto terzi, i brevetti, gli scavi archeologici o la gestione dei poli museali).

Una certa accortezza nel tener conto degli esiti della procedura è consigliabile in relazione a quanto reso esplicito nel rapporto finale, in particolare riguardo all'opportunità di non utilizzare i dati per confrontare i risultati della valutazione in aree scientifiche diverse (in ragione della peculiarità metodologica di ciascuna area), alla necessità di non prendere in considerazione la valutazione dei singoli ricercatori (dal momento che la procedura è stata declinata a livello di struttura e sottostruttura) e, infine, all'opportunità di non trasferire i risultati alla qualità dell'attività didattica svolta dalle università (che non è stata affatto presa in considerazione nella procedura).

2. Esiti nell'area 11 e nei settori scientifico-disciplinari pedagogici

Nell'ambito della procedura VQR, l'area 11 è stata suddivisa in due sub-aree: la 11a, composta dalle discipline di Antropologia, Filosofia, Geografia, Pedagogia, Scienze del Libro e del Documento, Storia, nell'ambito della quale è stata adottata una procedura di valutazione tramite *peer-review*, e la 11b, composta da Psicologia e Scienze motorie, nella quale si è fatto ricorso prevalentemente alle valutazioni bibliometriche (ANVUR, 2003b, p. 26).

Poiché i risultati di una valutazione tramite peer review non possono essere confrontati con gli esiti di procedure condotte con modalità bibliometriche, si fa qui riferimento ad alcuni esiti principali nell'area 11a, con

particolare riferimento ai settori scientifico-disciplinari pedagogici. Per un approfondimento sugli esiti nella sub-area 11b, si rinvia alla specifica parte del rapporto della VQR e, in particolare, alla tabella di distribuzione dei punteggi ottenuti per SSD nella sub-area bibliometrica (ANVUR, 2003d, tab. 2.23).

Va osservato che, nell'ambito della procedura, ciascun prodotto è stato collocato in una delle quattro classi di merito previste (Eccellenti, Buoni, Accettabili, Limitati), mentre quelli appartenenti a tipologie escluse dalla procedura o forniti con documentazione inadeguata sono stati classificati come «prodotti non valutabili». È stata infine prevista una penalizzazione per i casi accertati di plagio o frode [2].

Per ciò che concerne gli esiti principali della procedura in ambito pedagogico, va sottolineato, anzitutto, che si è registrata una grande partecipazione dei professori e ricercatori che hanno accettato di sottoporre gli esiti della propria attività di ricerca a valutazione. In linea con tutte le altre aree scientifiche, anche in ambito pedagogico la percentuale di prodotti conferiti rispetto a quella attesi è assai elevata: nel SSD M-PED/01 il 98.28% dei prodotti attesi sono stati conferiti, nel SSD M-PED/02 addirittura il 100%, nel SSD M-PED/03 il 99.26% e nel SSD M-PED/04 il 98.32 %. Con i loro 1625 prodotti conferiti (798 nel SSD M-PED/01, 248 nel SSD M-PED/02, 403 nel SSD M-PED/03, 176 nel SSD M-PED/04), tutti i settori pedagogici fanno così registrare percentuali di conferimento dei prodotti più alte della media della sub-area 11a (97.29%) (ANVUR, 2013d, tab. 2.1).

Dalla distribuzione dei prodotti si evince che in ambito pedagogico la tipologia più frequentemente conferita è quella della monografia, che costituisce rispettivamente il 44.2% dei prodotti per il SSD M-PED/01, il 38.7% dei prodotti per il SSD M-PED/02 e il 45.9% dei prodotti per il SSD M-PED/03. Fa eccezione il SSD M-PED/04 per il quale la tipologia di prodotto prevalente è il contributo in volume (35.2% dei prodotti presentati) (ANVUR, 2013d, tab. 1.2). Risulta così evidente come la monografia venga considerata uno strumento rilevante per la diffusione degli esiti della ricerca

[2] Si ricorda che le classi di merito e il relativo peso erano definite come segue. *Eccellente*: la pubblicazione si colloca nel 20% superiore della scala di valore condivisa dalla comunità scientifica internazionale (peso 1); *Buono*: la pubblicazione si colloca nel segmento 60%-80% (peso 0.8); *Accettabile*: la pubblicazione si colloca nel segmento 50%-60% (peso 0.5); *Limitato*: la pubblicazione si colloca nel 50% inferiore (peso 0); *Non valutabile*: la pubblicazione appartiene a tipologie escluse dal presente esercizio o presenta allegati e/o documentazione inadeguati per la valutazione o è stata pubblicata in anni precedenti o successivi al settennio di riferimento (peso -1); in casi accertati di *plagio o frode*, la pubblicazione è pesata con peso -2. Per ciascun prodotto mancante rispetto al numero atteso è stato assegnato un peso negativo pari a -0,5 (ANVUR, 2013b, p. 21).

in ambito pedagogico, non diversamente da quanto accade generalmente nel campo delle scienze umane.

La distribuzione dei prodotti di ricerca per lingua di pubblicazione mostra una limitata presenza di prodotti presentati in una lingua veicolare per la comunità scientifica di riferimento che li renda fruibili dalla maggior parte dei ricercatori stranieri potenzialmente interessati. I prodotti presentati in lingua italiana sono, infatti, il 78.57% del totale nel SSD M-PED/01, il 73.79% nel SSD M-PED/02, l'85.11% nel SSD M-PED/03, il 73.30% nel SSD M-PED/04: tutte percentuali al di sopra della media di sub-area (72.47%) (ANVUR, 2013d, tab. 1.4).

Una tale circostanza è probabilmente anche all'origine del dato relativo al ridotto numero di revisioni effettuate da referees stranieri nei settori pedagogici (194 in termini assoluti, l'11.94% del totale) (ANVUR, 2013d, tab. 2.5).

Dai giudizi assegnati attraverso la procedura peer review risulta un voto medio (indice I, con un valore compreso tra 0 e 1, calcolato sul numero complessivo di prodotti attesi) di 0.59 per i SSD M-PED/01 e M-PED/02, di 0.56 per i SSD M-PED/03 e di 0.58 per il SSD M-PED/04, con valori che si trovano al di sopra della media della sub-area (0.57) nel caso dei SSD M-PED/01, 02 e 04. Inoltre la distribuzione dei prodotti nelle classi di merito mostra che, in linea con quanto accade generalmente nell'intera sub-area, nei settori pedagogici le valutazioni si addensano nella fascia «Buono»: tale classe rappresenta il 46.92% delle valutazioni nel SSD M-PED/01, il 42.34% nel SSD M-PED/02, il 43.10% nel SSD M-PED/03 e il 44.13% nel SSD M-PED/04. Spicca la percentuale di giudizio di eccellenza ottenuto nel SSD M-PED-02 (19.76%) che risulta più alta di quella dell'area (ANVUR, 2013d, tab. 2.22, riprodotta qui in Tabella 1).

Se si analizza la distribuzione delle valutazioni per fascia di docenza si scopre che i giudizi di «eccellenza» sono stati assegnati più frequentemente alle fasce di professore ordinario per tutti i settori scientifico-disciplinari pedagogici (21.05% nella fascia P.O. contro 13.08% nell'intero settore M-PED/01; 37.21% contro 20.08% nell'intero settore M-PED/02; 25.49% contro 11.66% nell'intero settore M-PED/03; 14.58% contro 11.36% nell'intero settore M-PED/04) (ANVUR, 2013d, tab. 2.13).

Ulteriori esiti emersi dalla procedura VQR, che per brevità non vengono qui discussi in dettaglio, concernono la distribuzione dei prodotti giudicati «eccellenti» tra le strutture per i diversi settori scientifico-disciplinari (ANVUR, 2013d, tabb. 3.13, 3.14, 3.15 e 3.16) e la graduatoria delle strutture suddivise in «Grandi», «Medie» e «Piccole», ordinate per voto medio per i diversi settori scientifico-disciplinari (ANVUR, 2013d, tabb. 3.57, 3.58, 3.59 e 3.60).

Tabella 1. – Punteggi ottenuti e distribuzione dei prodotti nelle classi di merito per SSD, sub-area non bibliometrica (ANVUR 2013d, tab. 2.22).

SSD	Somma punteggi (V)	# Prodotti attesi (N)	Voto medio (I = V/N)	% Prodotti E	% Prodotti B	% Prodotti A	% Prodotti L	% Pprodotti penalizzati
M-DEA/01	253.00	467	0.54	13.06	37.47	24.41	23.13	1.93
M-FIL/01	248.70	449	0.55	12.03	42.09	21.83	21.83	2.23
M-FIL/02	150.40	254	0.59	23.62	38.58	14.17	19.69	3.94
M-FIL/03	358.80	558	0.64	19.00	49.46	17.38	9.14	5.02
M-FIL/04	161.80	245	0.66	24.08	43.27	18.37	12.24	2.04
M-FIL/05	163.60	291	0.56	15.46	40.21	18.90	23.71	1.72
M-FIL/06	446.20	714	0.62	21.29	46.78	13.03	14.29	4.62
M-FIL/07	93.60	141	0.66	31.91	33.33	15.60	19.15	0.00
M-FIL/08	87.90	114	0.77	44.74	37.72	7.02	7.89	2.63
M-GGR/01	278.80	551	0.51	6.17	41.92	23.96	25.77	2.18
M-GGR/02	177.60	357	0.50	9.24	34.17	30.25	23.25	3.08
M-PED/01	**481.80**	**812**	**0.59**	**12.81**	**46.92**	**20.44**	**17.73**	**2.09**
M-PED/02	**147.00**	**248**	**0.59**	**19.76**	**42.34**	**14.52**	**21.77**	**1.61**
M-PED/03	**227.50**	**406**	**0.56**	**11.58**	**43.10**	**20.69**	**23.89**	**0.74**
M-PED/04	**104.20**	**179**	**0.58**	**11.17**	**44.13**	**25.14**	**17.88**	**1.68**
M-STO/01	338.60	548	0.62	18.80	45.99	16.61	15.15	3.47
M-STO/02	501.70	823	0.61	17.50	43.62	22.11	12.27	4.50
M-STO/03	47.40	107	0.44	11.21	35.51	20.56	25.23	7.48
M-STO/04	611.30	1247	0.49	8.42	38.17	25.34	23.90	4.17
M-STO/05	103.70	173	0.60	19.65	42.77	16.18	18.50	2.89
M-STO/06	45.90	74	0.62	18.92	44.59	20.27	10.81	5.41
M-STO/07	138.90	245	0.57	16.33	50.20	11.43	13.47	8.57
M-STO/08	135.20	221	0.61	15.38	40.27	27.15	17.19	0.00
M-STO/09	120.90	184	0.66	32.61	39.67	14.13	3.80	9.78
n.a.	192.40	370	0.52	11.35	40.00	21.35	23.78	3.51
Totale	**5616.90**	**9778**	**0.57**	**15.42**	**42.37**	**20.31**	**18.52**	**3.37**

Punteggi ottenuti e distribuzione dei prodotti nelle classi di merito (Eccellente -E-, Buono -B-, Accettabile -A-, Limitato -L-) per SSD. Per «Somma punteggi (V)» si intende la valutazione complessiva del SSD ottenuta sommando i punteggi dei prodotti afferenti al SSD. La categoria «Prodotti penalizzati» contiene i prodotti non valutabili e casi accertati di plagio o frode così come previsto dal bando VQR del 7 novembre 2011, i prodotti mancanti (cioè attesi e non conferiti), i prodotti identici presentati più volte dalla stessa struttura, i prodotti identici presentati più volte dallo stesso soggetto valutato per due strutture di tipologia differente (es. università ed ente di ricerca). Per «# Prodotti attesi» si intende il numero di prodotti attesi calcolato sulla base del SSD di afferenza dei soggetti valutati e del numero di prodotti che da bando questi erano tenuti a inviare alla VQR.

Va osservato che le graduatorie delle strutture dipartimentali vengono presentate sia in relazione all'assetto precedente alla L. 240 (ANVUR, 2013d, tab 4.3), sia in relazione all'assetto post L. 240 (ANVUR, 2013d, tab 4.8), sia infine in base alle valutazioni ottenute dai quattro settori pedagogici (ANVUR, 2013d, tabb. 4.22, 4.23, 4.24 e 4.25).

Particolarmente interessante è, infine, la sezione del rapporto nella quale si formulano ipotesi per integrare i diversi indicatori di area in un indicatore finale di struttura per le università e per gli enti di ricerca in base a specifici pesi (ANVUR, 2013b, tabb. 6.10a, 6.10b, 6.11a, 6.11b, 6.12a e 6.12b).

3. Osservazioni conclusive

Dopo aver presentato un sintetico profilo delle modalità e degli esiti della procedura VQR, appare utile svolgere alcune riflessioni conclusive, che attengono, più in generale, il tema della valutazione della ricerca. Va in primo luogo ribadito che la valutazione costituisce una risorsa importante per il sistema universitario e della ricerca, in un'ottica di miglioramento progressivo della qualità e della competitività su scala internazionale: l'ampia partecipazione dei professori e ricercatori italiani, che hanno responsabilmente accettato di sottoporre a valutazione gli esiti delle propria attività di ricerca, rappresenta un segnale estremamente importante in un'ottica di *accountability* e per l'ulteriore diffusione di una cultura della valutazione, in grado di consentire la rilevazione di informazioni essenziali per il governo del sistema e per la formulazione di scelte consapevoli di indirizzo.

Riflettere criticamente sulle modalità adottate in un esercizio della valutazione così ampio come quello appena concluso è poi strategico anche in vista delle nuove procedure che potranno essere realizzate in futuro. In questo senso, un segnale importante nell'ambito dell'area 11 è stato espresso anche con la giornata di riflessione sulla VQR 2004-2010 organizzata dalle società scientifiche di tale area il 31 ottobre 2013 presso il Ministero dell'Istruzione dell'Università e della Ricerca, a cui hanno partecipato anche diversi rappresentanti del Consiglio Direttivo dell'ANVUR.

Si propongono di seguito alcuni elementi per una riflessione sulla procedura adottata nell'ambito della VQR. Un primo aspetto che occorre considerare è l'opportunità di una certa cautela, di certo non dimostrata dai mezzi di comunicazione di massa, nell'utilizzo delle graduatorie delle strutture. In molti casi, con scarti tra una posizione e l'altra riferibili a valori ridottissimi, l'errore statistico rende poco significativo attribuire posizioni differenziate in una graduatoria. A questo proposito il Consiglio Universitario Nazionale

(CUN), nella sua Dichiarazione del 16 luglio 2013, ricorda che, in un'ottica di promozione di occasioni di «crescita del sistema» e non di processi che favoriscano «divisioni, rivalità immotivate o logiche punitive, […] una valutazione rigorosa implica innanzitutto la capacità di fare e pubblicare confronti omogenei, senza centrare l'analisi su dati ad effetto». Ciò di cui si avverte il bisogno è, quindi, «valutazione e non classifiche. Rating e non ranking».

Ulteriore prudenza è opportuna in riferimento al calcolo e all'uso degli indicatori finali di qualità della ricerca contenuti nella VQR. Come auspica il CUN nella sua Raccomandazione del 9 ottobre 2013, il bando VQR 2004-2010 prevedeva la determinazione di sette indicatori di qualità (IRS1-IRS7) per le strutture e di quattro indicatori di qualità (IRD1-IRD4) per le sottostrutture (ovvero i dipartimenti), calcolati separatamente per le 14 aree scientifiche, nonché l'integrazione di tali indicatori in un unico indicatore di qualità finale della ricerca della struttura (IRFS) e in un unico indicatore di qualità finale della ricerca svolta dalla sottostruttura (IRFD), da utilizzare per le scelte di governo del sistema. Lo stesso rapporto finale della VQR segnala, a tal proposito, che la scelta delle modalità di integrazione degli indicatori calcolati per le aree in un unico indicatore (ovvero gli indicatori finali IRFS e IRFD) è di competenza del MIUR per le strutture e delle singole strutture per le sottostrutture: per questo motivo, quello contenuto nel rapporto finale della VQR rappresenta un mero esempio di calcolo degli indicatori finali IRFS e IRFD (ANVUR, 2013e, p. 1). Come osserva il CUN, alla luce di tutto ciò, appare doveroso evitare di applicare in modo automatico i risultati degli indicatori della VQR, effettuando invece il calcolo e l'uso degli indicatori finali «con procedure che rendano chiaro il significato degli indicatori utilizzati e che dipendano dallo scopo previsto per l'uso di tali indicatori».

In vista di un nuovo esercizio della valutazione, che senza dubbio dovrebbe essere condotto con cadenza regolare nel sistema universitario, occorre allora avviare un serio dibattito sul tema della valutazione della ricerca scientifica, anche in un'ottica di progressivo miglioramento rispetto a quanto fatto in passato. Nella consapevolezza che è importante coinvolgere nella discussione anche chi si è occupato nella propria attività di ricerca di metodologia della valutazione (e non è il caso di chi scrive), si segnalano, a conclusione del presente contributo, alcuni primi interrogativi sui quali il dibattito dovrebbe concentrarsi:

- Considerata la complessità del panorama della ricerca, quali modalità sono più opportune per valutare l'attività di ricerca in ambito umanistico?
- Quali esperienze di esercizi della valutazione in ambito internazionale possono offrire utili riferimenti per il contesto italiano?
- Quali modalità della valutazione possono essere adottate per evitare di fiaccare ulteriormente un sistema universitario come quello italiano di fatto indebolito da un perdurante stato di sottofinanziamento?

- Quale modalità della valutazione potrebbe tener conto degli esiti conseguiti nelle strutture e nelle sottostrutture anche in relazione ai finanziamenti precedentemente ottenuti da tali organismi per la loro attività di ricerca?
- Quali modalità della valutazione è possibile adottare per la definizione di procedure che siano effettivamente sostenibili in termini di costi, al fine di dar vita a procedure di valutazione stabili e continuative?
- Quali modalità di autovalutazione è possibile affiancare ai processi di valutazione finora condotti, al fine di promuovere processi partecipativi di riflessione all'interno delle strutture e delle sottostrutture?

Riferimenti bibliografici

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013a). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale. 30 Giugno 2013*. http://www.anvur.org/rapporto/ (consultato il 21/11/2013).

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013b). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale ANVUR. Parte Prima: Statistiche e risultati di compendio. 30 Giugno 2013*. http://www.anvur.org/rapporto/files/VQR2004-2010_RapportoFinale_parteprima.pdf (consultato il 21/11/2013).

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013c). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area 11 (GEV 11)*. http://www.anvur.org/rapporto/files/Area11/VQR2004-2010_Area11_RapportoFinale.pdf (consultato il 21/11/2013).

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013d). *Tabelle area 11 in pdf*. http://www.anvur.org/rapporto/files/Area11/VQR2004-2010_Area11_Tabelle.pdf (consultato il 21/11/2013).

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013e). *Appendice D. Il calcolo dei pesi per la composizione delle valutazioni di Area*. http://www.anvur.org/rapporto/files/Appendici/VQR2004-2010_AppendiceD.pdf (consultato il 21/11/2013).

Alcune considerazioni statistiche sulla VQR 2004-2010

Rosa Capobianco

Università degli Studi «Roma Tre», Dipartimento di Scienze della Formazione

rcapobianco@uniroma3.it

1. Introduzione

La Valutazione della Qualità della Ricerca (VQR), relativa al periodo 2004-2010, ha riguardato 14 aree disciplinari. I rapporti finali di area sono molto dettagliati ed esaustivi, ed illustrano in maniera articolata sia le fasi della valutazione sia i punti di forza che le criticità emerse. In questa breve nota si riportano alcune considerazioni relative alla valutazione dei lavori scientifici che hanno portato alla successiva graduatoria delle strutture universitarie, con particolare riferimento al rapporto finale dell'area 11 «Scienze storiche, filosofiche, pedagogiche e psicologiche».

L'area 11 è una realtà molto composita, formata da 38 settori scientifico-disciplinari (raggruppati in 8 macro-settori), molto eterogenei tra loro sia per tradizione culturale e tipologia dei lavori scientifici – per cui si è resa necessaria la distinzione dell'area 11 nella sub-area non bibliometrica e nella sub-area bibliometrica – sia per numero di docenti afferenti e relativa presenza nelle università, come mostrato dal grafico di Figura 1 in cui è riportato il numero di Atenei in cui ciascun macro-settore è presente.

È opportuno precisare che le elaborazioni realizzate in questo lavoro fanno riferimento ad alcune delle tabelle pubblicate [1] sul sito dell'ANVUR, richiamate nel testo, che per motivi di brevità non sono state riportate. Nelle tabelle consultate non sono riportate le strutture con meno di 10 lavori attesi, per questo motivo alcune analisi potrebbero discostarsi dai risultati ufficiali della VQR.

[1] Le elaborazioni fanno riferimento alle tabelle 2.15, 2.16, 2.17, 2.18, 3.37, 3.38, 3.41 e 3.48, pubblicate sul sito dell'ANVUR, nella sezione dedicata al rapporto finale dell'area 11.

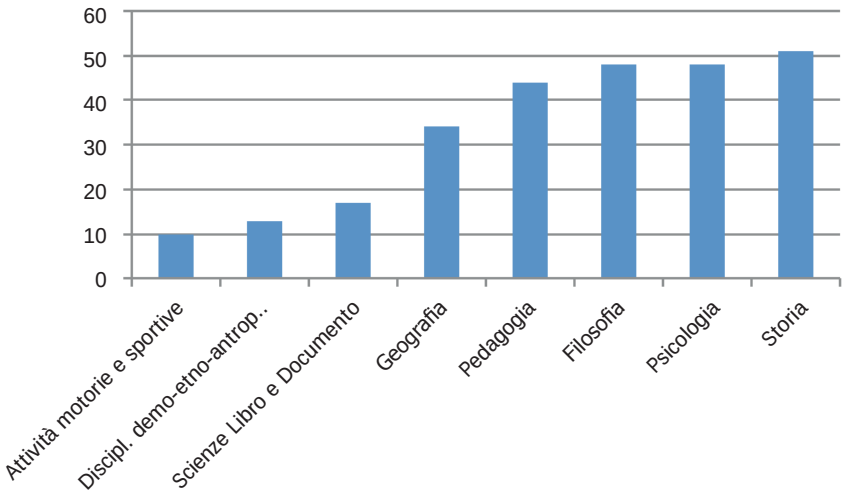


Figura 1. – Distribuzione dei macro-settori dell'area 11 per numero di Atenei.

2. La valutazione dei prodotti della ricerca

Come previsto dal bando ministeriale, i prodotti della ricerca sono stati giudicati Eccellenti (E), Buoni (B), Accettabili (A) e Limitati (L) con punteggi 1, 0.8, 0.5 e 0. Ai prodotti mancanti (MIS) è stato assegnato peso -0.5, ai prodotti non valutabili (NV) peso -1 e ai casi accertati di plagio (PL) peso -2. Tali valutazioni non sono tra loro equidistanti [2] e danno maggiore risalto alle valutazioni negative piuttosto che a quelle positive: il campo di variazione dei prodotti giudicati positivamente, infatti, è [0, 1] notevolmente più piccolo rispetto al campo di variazione delle valutazioni negative, compreso tra [-2, 0]. Una trasformazione dei giudizi più equilibrata potrebbe essere quella illustrata in Tabella 1.

Come è facile dedurre dalla tabella, in seguito a questa attribuzione di valori, il campo di variazione dei prodotti è tra -1 e 1. Sarebbe interessante poter ricalcolare le valutazioni sulla base di questa nuova griglia, tuttavia dai dati aggregati presenti sul sito dell'ANVUR è possibile conoscere esclusivamente le percentuali di lavori giudicati Eccellenti, Buoni, Accettabili e Limitati, mentre sono riportate aggregate le percentuali di prodotti penalizzati (mancanti, non valutabili e plagio). Le tabelle 2.1 e 2.20 riportano infatti, per

[2] K. D. Bailey, *Metodi della Ricerca Sociale*, Bologna: Il Mulino, 1994, p. 405.

la sub area non bibliometrica, rispettivamente la percentuale di prodotti mancanti e la percentuale di prodotti penalizzati per ogni settore scientifico disciplinare, così come le tabelle 2.2 e 2.21 per la sub area bibliometrica; mentre le tabelle 3.37 e 3.38 riportano la percentuale di prodotti penalizzati per ogni Ateneo sia per la sub-area non bibliometrica che per quella bibliometrica.

Tabella 1. – Trasformazione dei giudizi.

Giudizio	Valore
Eccellente	1
Buono	0,75
Accettabile	0,50
Limitato	0,25
Non valutabile	0
Mancante	-0,5
Plagio	-1

3. Gli indicatori di qualità delle strutture

Sulla base della griglia di valutazione predisposta dall'ANVUR sono stati costruiti gli indicatori di qualità sia delle aree tematiche che dei dipartimenti e degli Atenei. È stato definito l'indicatore della struttura i-esima nell'area j-esima (j = 1 …, 14),

$$I_{i,j} = \frac{v_{i,j}}{n_{i,j}} \tag{1}$$

dove $v_{i,j}$ è pari a

$$v_{i,j} = n_{i,j}E + 0.8n_{i,j}B + 0.5n_{i,j}A + 0n_{i,j}L - 0.5n_{i,j}MIS - n_{i,j}NV - 2n_{i,j}PL$$

mentre $n_{i,j}$ rappresenta il numero di lavori attesi per la struttura i-esima nell'area j-esima. L'indicatore $I_{i,j}$ varia tra [-2, 1]: assume valore -2 nel caso in cui tutti i lavori presentati dalla struttura i-esima nell'area j-esima siano stati classificati come «Plagio», mentre assume valore 1 se tutti i lavori sono stati classificati «Eccellenti».

Per effetto dei valori attribuiti ai prodotti di ricerca, l'indicatore (1) è molto sensibile a variazioni millesimali. Osservando, ad esempio, la graduatoria riportata nella tab. 3.38, relativa all'sub-area bibliometrica dell'area 11,

è facile notare come il posto occupato in graduatoria dalle strutture varia per pochi millesimi. L'Università «Roma Tre», ad esempio, risulta al 40-esimo posto in graduatoria con un indice pari a 0,3429, compresa tra l'Università di Enna (0,3281) e l'Università Napoli «Parthenope» (0,3432). In questa sub-area l'Università «Roma Tre» ha complessivamente 28 lavori attesi che hanno riportato il punteggio 9,6, da cui I = 9,6/28 = 0,3429. Ammettendo, per ipotesi, un errore involontario nella valutazione dei lavori dell'Università «Roma Tre», per cui ad un lavoro giudicato accettabile (0,5) è stato invece attribuito il valore 0 (giudizio limitato), la somma dei punteggi risulterebbe pari a 10,1 e – a parità di lavori attesi – l'indicatore sarebbe uguale a I = 10,1/28 = 0,3607 con una conseguente variazione della graduatoria per cui l'Università «Roma Tre» passerebbe dal 40-esimo al 37-esimo posto. È opportuno precisare che la sensibilità dell'indicatore è inversamente proporzionale al numero di lavori attesi: più piccolo è $n_{i,j}$, maggiore è l'effetto sull'indice causato da una variazione della valutazione.

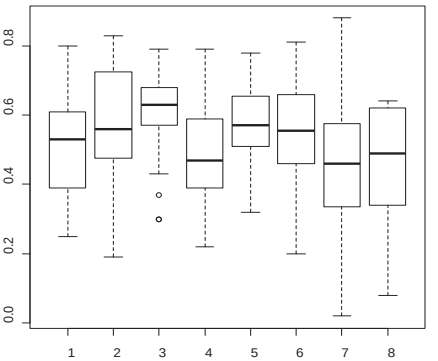
L'esempio precedente non è soltanto un esercizio statistico ma un'eventualità abbastanza plausibile visto che nell'area 11 le percentuali di valutazioni concordanti sono abbastanza basse: solo il 37,18%, nella sub-area non bibliometrica (vd. tab. 2.17) e il 20,72% nella sub-area bibliometrica (vd. tab. 2.18). In particolare, come risulta dalle tabelle 2.15 e 2.16, i settori M-PED01-03-04 sono tra quelli che hanno avuto la percentuale di valutazioni discordanti (Eccellente *vs* Limitato) più alte nell'area [3]: in ordine decrescente i settori M-PED04, M-PED01 e M-PED03 occupano le posizioni 3, 4 e 6. In ogni caso è abbastanza bassa (una media dell'area pari al 77%) la percentuale di lavori su cui si è arrivati ad una valutazione concordante o minimamente discordante.

Rappresentando le distribuzioni dell'indicatore (1) per gli 8 macro-settori [4], desunte dalle tabelle 3.41-3.48 (vd. Figura 2), si può notare che le distribuzioni presentano valori centrali diversi che variano tra 0,46 (Psicologia) e 0,63 (Filosofia), eteroschedasticità e nella maggior parte dei casi asimmetria negativa. Nel caso della distribuzione dell'indicatore per il macro-settore di Filosofia sono presenti anche due valori anomali.

Il grafico evidenzia in modo più incisivo quanto osservato a proposito del campo di variazione dell'indicatore $I_{i,j}$ che per valutazioni positive è, teoricamente, tra [0, 1] mentre i dati reali mostrano che è in realtà più stretto (vd. Tabella 2)

[3] Il settore M-PED04 ha avuto una percentuale di valutazioni discordanti pari al 5,81%, il settore M-PED01 il 5,68% e il settore M-PED03 il 5,37%.

[4] I dati rappresentati fanno riferimento alle tabelle presentate sul sito dell'ANVUR in cui non compaiono le strutture con meno di 10 prodotti attesi.



1 = Discipline demo-etno-antropologiche; 2 = Scienze del Libro e del Documento, 3 = Filosofia; 4 = Geografia; 5 = Pedagogia; 6 = Storia; 7 = Psicologia; 8 = Attività motorie e sportive.

Figura 2. – Box-plot in parallelo relativi alle distribuzioni dell'indicatore $I_{i,j}$ per gli 8 macro-settori.

Tabella 2. – Valore minimo, massimo e campo di variazione per gli 8 macro-settori dell'area 11.

Macro-settore	Valore minimo	Valore massimo	Campo di variazione
Discipline demo-etno-antropologiche	0,25	0,80	0,55
Scienze del Libro e del Documento	0,19	0,85	0,66
Filosofia	0,30	0,79	0,49
Geografia	0,22	0,79	0,57
Pedagogia	0,32	0,78	0,46
Storia	0,20	0,81	0,61
Psicologia	0,02	0,88	0,86
Attività motorie e sportive	0,08	0,64	0,56

Le motivazioni per queste diminuzioni del campo di variazione sono molteplici e ben evidenziate nel rapporto finale in cui tra l'altro si sostiene «[…] la peer review è infatti un sistema soggettivo, ed è molto difficile che i referee assegnino alle fasce estreme, e in particolare a quelle dell'eccellenza più del 10-15% dei prodotti» [5], o come dimostrato dal grafico di Figura 3 da cui si evince che i lavori presentati su riviste di classe A sono stati giudicati prevalentemente (50% circa dei casi) «Buoni».

[5] ANVUR, *Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area 11 (GEV 11)*, 2013, p. 25.

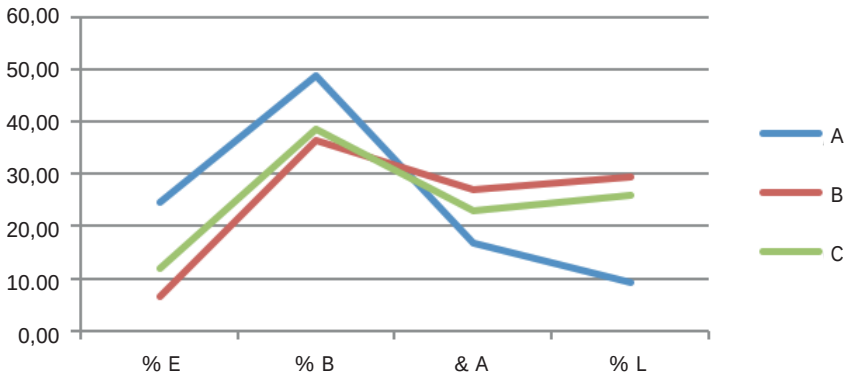


Figura 3. – Distribuzione delle valutazioni per fasce di classificazione delle riviste.

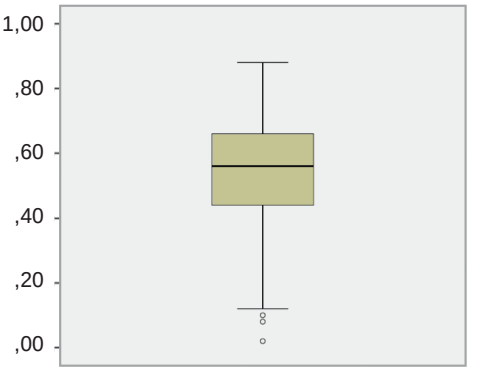


Figura 4. – Distribuzione congiunta degli indicatori $I_{i,j}$ per l'area 11.

A partire dall'indicatore $I_{i,j}$ sono stati definiti gli altri indicatori. Ad esempio le distribuzioni dell'indicatore $I_{i,j}$, così eterogenee tra loro, sono state accorpate per definire l'indicatore $R_{i,j}$ dato dal rapporto

$$R_{i,j} = \frac{I_{i,j}}{V_{i,j}/N_j}, \text{ per j = 1, …, 14}$$

dove $V_j = \sum_{i,1}^{N_{ST}} v_{i,j}$ è la valutazione complessiva dell'area j-esima mentre $N_j = \sum_{i=1}^{N_{ST}} n_{i,j}$ è il numero totale di prodotti attesi per l'area j-esima, N_{ST} rappresenta il numero totale di strutture presenti nell'area 11. L'indicatore $R_{i,j}$ è il

rapporto tra il voto medio ricevuto dai prodotti della struttura i-esima nell'area j-esima e il voto medio ricevuto da tutti i prodotti dell'area j-esima. Dal momento che le distribuzioni non sono né omoschedastiche né «di forma normale», accorpare tutti i valori $I_{i,j}$ in un'unica distribuzione può dare origine ad una distribuzione «non normale» (vd. Figura 4) e pertanto il confronto dei valori $I_{i,j}$ con la media totale può portare ad avere risultati poco affidabili. Nel caso in esame, vista anche la presenza di valori anomali nella distribuzione, sarebbe stato opportuno usare indici meno sensibili alla presenza di *outlier* o a deviazioni dalla normalità, quali la mediana o la media troncata, piuttosto che la media aritmetica.

Concludiamo questa breve nota con una considerazione relativa alla suddivisione delle strutture in «Grandi», «Medie» e «Piccole», qualora ci sia un elevato numero di docenti afferenti al settore scientifico disciplinare. Nella tabella 3.37, per la sub-area non bibliometrica, le strutture sono state suddivise in base al numero di prodotti attesi: piccole ($n < 100$), medie ($100 \leq n \leq 299$) e grandi ($n \leq 300$). Anche per la sub-area bibliometrica è stata fatta la stessa ripartizione ma usando valori soglia diversi: piccole ($n < 30$), medie ($30 \leq n \leq 99$) e grandi ($n \leq 100$) (vd. tab. 3.38). Ovviamente la determinazione dei valori soglia è fondamentale perché da essa dipende la graduatoria delle strutture ma, in entrambi i casi, non è chiaro il criterio in base al quale i valori sono stati determinati. In questo caso non è stato possibile fare simulazioni poiché nelle tabelle non compaiono le strutture che hanno meno di 10 lavori attesi.

I parametri della VQR 2004-2010 e il controllo della soggettività nelle procedure di valutazione

Valeria Biasi

Università degli Studi «Roma Tre», Dipartimento di Scienze della Formazione
Membro del Nucleo di Valutazione dell'Ateneo «Roma Tre»

valeria.biasci@uniroma3.it

1. Principali criteri adottati per la procedura di valutazione della qualità della ricerca VQR

Il processo di valutazione VQR ha comportato l'applicazione di specifici criteri ed indicatori per valutare, da una parte, i *prodotti della ricerca* e, dall'altra, la qualità delle *strutture di ricerca*.

Per quanto riguarda i prodotti di ricerca (articoli, monografie, brevetti, software, ecc.) – valutati dai GEV (Gruppo di Esperti della Valutazione) – sono stai adottati i criteri di rilevanza, originalità e grado di internazionalizzazione con due diverse metodologie: valutazione bibliometrica e *peer-review*.

Così facendo all'ANVUR sono pervenuti in totale 184.878 prodotti, contro i 194.763 attesi. La percentuale media sulle aree di prodotti mancanti è stata del 5,1%. I GEV per valutare i lavori potevano, in base alle indicazioni fornite dalle Società scientifiche e alle mediane del Settore scientifico-disciplinare, utilizzare l'analisi bibliometrica basata sul numero di citazioni del prodotto e sull'*Impact Factor* della rivista dove questo era pubblicato; oppure la procedura *peer-review* affidando a *referees*, scelti appunto dal GEV (di norma due per prodotto). Complessivamente sono stati impegnati nella procedura *peer-review*, 14.770 revisori «attivi» (che cioè hanno svolto almeno una revisione), di cui 10.150 italiani e 4.620 con affiliazione estera.

In genere i criteri utilizzati hanno portato ad una sorta di penalizzazione di quelle strutture con significativa presenza di docenti impegnati più nella didattica e nell'organizzazione rispetto alla ricerca; ma proprio quando si intende valutare la qualità della ricerca scientifica, come si può «dimenticare» il criterio fondamentale a garanzia della scientificità del prodotto, ossia il rigore metodologico, la correttezza e l'appropriatezza dei metodi adottati

per sostenere le conclusioni raggiunte? Una carenza a nostro avviso da sanare in futuro.

Esaminando la tipologia dei prodotti, risulta che per le aree delle Scienze matematiche e informatiche (area 1), Scienze biologiche (area 5) e per le Scienze economiche e statistiche (area 13), gli articoli su rivista costituiscono la stragrande maggioranza dei prodotti presentati. Nelle aree 10, 11, 12, invece, le monografie sono in maggioranza. Gli esiti principali nella sub-area 11a, per la *qualità scientifica dei prodotti* presentati, con particolare riferimento ai settori scientifico-disciplinari pedagogici, ottenuti tramite la valutazione *peer-review* (ANVUR, 2013d, tab. 2.22) sono stati esposti in modo dettagliato nell'intervento curato da Marco Catarci in apertura del presente dibattito.

Gli esiti principali ottenuti nell'area 11b – sub-area bibliometrica – sono riassunti nella seguente tabella di distribuzione dei punteggi (ANVUR, 2013d, tab. 2.23), che presentiamo e commentiamo in Tabella 1.

Si rileva una complessiva elevata produttività della sub-area 11b, la quale non appare tuttavia esente da aspetti problematici in particolare nell'ambito del confronto tra i diversi settori psicologici. Come si evince la distribuzione dei prodotti nelle classi di merito mostra appunto che, in questi settori, le valutazioni si addensano nella fascia «Eccellente»: in particolare nel SSD M-PSI/02, addirittura con l'84.70%, seguito a notevole distanza dal SSD M-PSI/01, con il 45.99%, e con punteggi a scalare per altri settori SSD psicologici. Come avremo modo di spiegare nel paragrafo 2.1, la distribuzione della frequenza delle riviste con *Impact Factor* non è omogenea nei vari settori scientifici e, non solo, i coefficienti di IF più elevati si registrano per riviste dedicate a tematiche fisiologiche, neurofisiologiche, mediche, biologiche, ecc., per cui, non tenendo conto in modo proporzionale di questo dato di realtà, si originano delle distorsioni inevitabili sulla validità dei prodotti di ricerca. Vale a dire, nel caso specifico illustrato nella Tabella suddetta, è il SSD M-PSI/02 della «Psicobiologia e Psicologia fisiologica» a fare la parte del leone: cioè il settore che ha a disposizione il numero maggiore di riviste con alto IF, non paragonabile all'offerta disponibile per gli altri settori. Ciò introduce un criterio di iniquità molto pesante, e – in particolare per il tipo di analisi che qui intendiamo tracciare – introduce una *distorsione metodologica*: la comparazione di insiemi di dati non omogenei, e su questo sarà utile orientare il dibattito futuro in tema di valutazione della qualità della ricerca.

Sul fronte della lingua di pubblicazione (considerata un indicatore del grado d'internazionalizzazione), il 62,1% dei prodotti è in inglese con notevoli differenze tra le aree: per esempio abbiamo l'88,6% in Ingegneria industriale e dell'Informazione e il 5,7% in Scienze giuridiche; nell'area delle Scienze dell'antichità, filologico-letterarie e storico-artistiche il 13,2% dei prodotti conferiti sono in una lingua straniera diversa dall'inglese.

Tabella 1. – Punteggi ottenuti e distribuzione dei prodotti nelle classi di merito per SSD, sub-area bibliometrica (ANVUR 2013d, tab. 2.23).

SSD	Somma punteggi (V)	# Prodotti attesi (N)	Voto medio (I = V/N)	% Prodotti E	% Prodotti B	% Prodotti A	% Prodotti L	% Prodotti penalizzati
M-EDF/01	80.70	168	0.48	26.79	23.21	7.74	39.29	2.98
M-EDF/02	87.45	168	0.52	26.79	29.17	7.74	33.93	2.38
M-PSI/01	**471.40**	**761**	**0.62**	**45.99**	**11.56**	**14.72**	**25.89**	**1.84**
M-PSI/02	**323.90**	**353**	**0.92**	**84.70**	**7.93**	**1.98**	**5.10**	**0.28**
M-PSI/03	**110.10**	**194**	**0.57**	**35.05**	**16.49**	**20.10**	**26.80**	**1.55**
M-PSI/04	**252.90**	**500**	**0.51**	**24.80**	**13.60**	**30.40**	**30.80**	**0.40**
M-PSI/05	**233.90**	**417**	**0.56**	**34.29**	**12.71**	**25.42**	**25.66**	**1.92**
M-PSI/06	**97.30**	**209**	**0.47**	**17.70**	**10.05**	**43.06**	**28.23**	**0.96**
M-PSI/07	**135.30**	**322**	**0.42**	**7.45**	**9.63**	**54.35**	**27.95**	**0.62**
M-PSI/08	**181.70**	**460**	**0.40**	**23.70**	**10.65**	**23.04**	**35.87**	**6.74**
n.a.	60.20	157	0.38	29.94	21.66	9.55	17.20	21.66
Totale	**2034.85**	**3709**	**0.55**	**34.81**	**13.27**	**22.32**	**26.75**	**2.86**

Punteggi ottenuti e distribuzione dei prodotti nelle classi di merito (Eccellente -E-, Buono -B-, Accettabile - A-, Limitato -L-) per SSD. Per «Somma punteggi (V)» si intende la valutazione complessiva del SSD ottenuta sommando i punteggi dei prodotti afferenti al SSD. La categoria «Prodotti penalizzati» contiene i prodotti non valutabili e casi accertati di plagio o frode così come previsto dal bando VQR del 7 novembre 2011, i prodotti mancanti (cioè attesi e non conferiti), i prodotti identici presentati più volte dalla stessa struttura, i prodotti identici presentati più volte dallo stesso soggetto valutato per due strutture di tipologia differente (es. università ed ente di ricerca). Per «# Prodotti attesi» si intende il numero di prodotti attesi calcolato sulla base del SSD di afferenza dei soggetti valutati e e del numero di prodotti che da bando questi erano tenuti a inviare alla VQR.

Il grado d'internazionalizzazione nelle Scienze mediche è pari all'81.3%. Risulta, inoltre, consolidata l'abitudine a pubblicare i risultati della ricerca medica e biologica su riviste censite da banche dati quali *Web of Science* (WoS) di *Thomson Reuters* o *Scopus* poiché quasi il 90% dei prodotti ha avuto una valutazione bibliometrica.

La scala dei voti per i prodotti valutati va da zero, che equivale a un livello «limitato», a uno, che rappresenta l'eccellenza. In particolare i prodotti della ricerca sono stati giudicati Eccellenti (E), Buoni (B), Accettabili (A) e Limitati (L) con punteggi 1, 0.8, 0.5 e 0. Ai prodotti mancanti (MIS) è stato assegnato peso -0.5, ai prodotti non valutabili (NV) peso -1 e ai casi accertati di plagio (PL) peso -2. Tali valutazioni non sono tra loro equidistanti e di conseguenza vi è un campo di variazione non omogeneo, che co-determinerà un effetto negativo sulla stabilità dei dati finali, come dimostrato in modo raffinato dall'intervento curato da Rosa Capobianco, sempre in apertura del presente dibattito.

Sulla base dell'attribuzione di tali votazioni emergono delle graduatorie. Nella graduatoria delle eccellenze spiccano così le Scienze chimiche (0,78) e quelle fisiche, seguite da Ingegneria industriale e dell'Informazione e Scienze dell'antichità, filologico-letterarie e storico-artistiche (0,66). Distanti dalle prime posizioni della classifica delle eccellenze ci sono le Scienze economiche e statistiche (0,32) e le Scienze sociali e politiche (0,45). Se prendiamo in considerazione la percentuale dei prodotti eccellenti, l'area dell'Ingegneria civile ne ha solo 8,9%. In vetta alla graduatoria ci sono sempre la Chimica e la Fisica, seguita dall'Ingegneria industriale e dell'informazione (53,82%), Architettura (42,03%), Scienze matematiche e informatiche (41,94%), Scienze biologiche (40,06%), Scienze della Terra (34,7%), Scienze mediche (33,96%), Scienze psicologiche (33,91%).

Una questione classificatoria si pone in modo ineludibile: a cosa serve, e se è poi deontologicamente corretto, o se abbia senso per il riconoscimento del grado di scientificità dei diversi saperi, ottenere una cosiddetta «Graduatoria per aree». Questa graduatoria, basata sulla produttività media delle singole aree, deve essere ben tenuta lontana da pregiudizi di superiorità di una scienza rispetto ad un'altra, pericolo che potrebbe negativamente orientare, secondo una lettura distorta.

Le università sono state inoltre divise in «Grandi», «Medie» e «Piccole» sulla base del numero dei soggetti valutati di ciascuna struttura e quindi il calcolo della produttività è stato visto in proporzione. Naturalmente la lettura va nel senso di rilevare quali strutture siano risultate ai primi posti, quali in pratica siano i migliori Atenei, nel rispetto delle grandi, medie e piccole dimensioni, quindi facendo confronti dentro gruppi per così dire omogenei almeno per numerosità di operatori. Queste informazioni possono essere uti-

li per la scelta universitaria da parte di studenti e famiglie. Un uso distorto di tali informazioni potrebbe invece orientare la maggior parte dei fondi ministeriali solo verso le eccellenze, a detrimento delle sedi che più necessitano di intervento e promozione.

Vale a dire che *gli esiti della valutazione sono utili quando, evidenziando punti critici, riescono a dare avvio a processi di intervento costruttivo*.

Per quanto concerne la cosiddetta «Terza missione», parametro che indica l'insieme delle attività con le quali i nostri Istituti di ricerca affiancano alle tradizionali missioni di insegnamento e di ricerca per operare verso una valorizzazione economica della conoscenza, come per esempio nel caso della produzione di Brevetti: le nostre università in media registrano un brevetto ogni 19 docenti, mentre gli enti di ricerca si attestano su 1 brevetto ogni 26. Naturalmente prendendo in esame la specificità delle aree disciplinari possiamo notare una distribuzione molto diversificata dei brevetti a scapito delle aree cosiddette «umanistiche», per le quali si dovrebbero forse prevedere altri criteri per declinare un trasferimento applicato alla realtà sociale.

In vista anche della prossima valutazione, che dovrebbe svolgersi entro cinque anni, occorre perciò sottolineare la necessità di introdurre dei «correttivi» per ridurre le principali «distorsioni» emerse dai risultati ottenuti.

2. Miti e limiti delle metodologie in uso per la valutazione della ricerca scientifica

La valutazione della produttività scientifica di università ed altri enti pubblici di ricerca, ai fini di una «oculata» distribuzione delle risorse, ha riscosso grande interesse in ambiente accademico sia in occasione della pubblicazione degli esiti iniziali del primo processo di Valutazione della Ricerca ad opera del CIVR sia, oggi, con i risultati della VQR.

Questioni deontologiche, epistemologiche, metodologiche hanno fatto da sfondo a scenari politici di ampio respiro che potremmo definire di economia politica della cultura.

La responsabilità degli enti addetti alla valutazione appare molto forte: la scelta della *metodologia valutativa*, l'appropriatezza e l'efficacia degli *indicatori* utilizzati, la *significatività* dei risultati, ricadono direttamente sul processo di distribuzione delle risorse finanziarie ed indirettamente sul riconoscimento delle aree scientifiche e sul peso politico di alcune strutture di ricerca rispetto ad altre.

È da segnalare che dopo la prima rilevazione ministeriale (Valutazione Triennale della Ricerca o VTR), i cui risultati parziali riguardanti il triennio

2001-2003 sono stati diffusi dai componenti del CIVR, in particolare dal presidente prof. Franco Cuccurullo (2007), il processo di valutazione si interruppe inaspettatamente o forse precauzionalmente. Vi fu la espressa richiesta del CUN, datata 7 maggio 2009, inoltrata all'allora Ministro, di riprendere tale attività, in particolare tenendo conto non solo del sistema di valutazione *peer-review* ma anche degli *indicatori bibliometrici* diffusi a livello internazionale. Si auspicò quindi che si potesse procedere, in base a tali indicatori, alla valutazione quinquennale (forse il sistema ministeriale di valutazione riecheggiò nella nostra mente un riferimento non casuale ai «piani» quinquennali di sovietica memoria) ed alla conseguente e rapida distribuzione delle ricchezze su tali basi.

Tutto ciò senza tenere ragionevole conto degli innumerevoli rilievi critici conseguenti alla prima valutazione su scala nazionale ha prodotto (cfr. Domenici, 2004); e quindi senza una riflessione costruttiva sull'*affinamento degli indicatori* da utilizzare per l'immediato futuro.

Quali correttivi si sarebbero dovuti apportare rispetto alla prima valutazione nazionale, inevitabilmente sperimentale? Passare da una valutazione triennale ad una quinquennale avrebbe potuto aiutarci a fare di necessità virtù, ma le serie di risultati non sarebbero state più confrontabili per aggiustamenti metodologici. Che fare? Siamo poi passati ad una valutazione che ha coperto un settennio, il 2004-2010, forse in un'ottica di sanatoria, per coprire incredibili lungaggini ed ora disponiamo da pochi mesi dei nuovi dati.

Il problema generale è ancora l'interpretazione e l'uso che sarà fatto di tali informazioni: ricorrere a criteri meritocratici per l'assegnazione delle risorse obbliga gli esperti di settore allo elaborazione di *criteri ottimali per la valutazione* ed allo sviluppo di metodologie efficaci in termini di *affidabilità* e *validità statistica*. Per quanto riguarda la validità si tratta di allestire strumenti di misura che rilevino con appropriatezza ciò che si intende misurare (*validità di costrutto*). Per quanto riguarda l'*affidabilità* degli strumenti adottati (la sensibilità del «termometro» utilizzato) occorre testare la *capacità di «fare lo zero»* (assenza di affidabilità del criterio) e via via le varie unità di misura della scala applicata.

Sappiamo certo che le più note metodologie per la valutazione della ricerca scientifica rientrano in due tipologie: la valutazione bibliometrica e la valutazione *peer-review*.

Cerchiamo di vedere sinteticamente quale validità e quale affidabilità presentano queste due distinte procedure di valutazione, e quindi inevitabilmente quali limiti: da tenere in debito conto per mantenere un atteggiamento valutativo equilibrato e non estremista.

2.1. *Metodi bibliometrici: vantaggi e limiti. Il mito dell'Impact Factor*

I metodi bibliometrici intendono rispondere ad un criterio quantitativo nel processo valutativo, sono collocati nel contesto della *scientometria*, definita come la «scienza per la misura e l'analisi della scienza». La bibliometria intende, attraverso tecniche matematiche e statistiche, analizzare la distribuzione delle pubblicazioni ed elaborare modelli per calcolare la ricaduta (o impatto) di questi prodotti sulla comunità scientifica.

La letteratura specialistica pone tuttavia spesso in evidenza la scarsa o nulla affidabilità del criterio bibliometrico nella *valutazione dell'originalità e del valore innovativo* del prodotto scientifico. Già MacRoberts e MacRoberts nel 1996 indicavano i problemi connessi ad un sistema di valutazione legato all'*indice delle citazioni*: *scarsa rappresentatività per i giovani ricercatori* non ancora «noti» nell'ambiente accademico, *insensibilità* (e conseguente affidabilità nulla per questo indicatore) nel distinguere le citazioni «positive» da quelle «critiche o negative» e quindi *assenza di capacità discriminativa*. Van Raan (2005) scrive proprio sulla rivista *Scientometrics* (n. 62, 1, pp. 133-143) sulla necessità di migliorare l'accuratezza del sistema bibliometrico.

L'analisi bibliometrica, usata in modo non sempre diffuso a livello internazionale e sviluppatasi negli ultimi vent'anni, è resa possibile dalla creazione delle banche-dati computerizzate, le quali non sono immuni da imprecisioni: autori non riconosciuti in quanto segnalati con affiliazioni universitarie diverse; processi di *ranking* effettuati più facilmente sui primi autori di articoli che sui co-autori, con possibilità di errore stimato pari al 30%.

Tra le varie misure bibliometriche, le due più conosciute sono il *numero di citazioni* ed il *fattore di impatto* o *Impact Factor*, corrispondente al numero di citazioni ricevute nell'anno in corso rispetto agli articoli pubblicati nei due anni precedenti, diviso per il totale del numero di articoli pubblicati negli stessi anni. Vi sono poi altri indici quali il *Cited Half Life* che misura il ciclo di vita di un lavoro scientifico, la durata nel tempo degli articoli citati o la durata delle citazioni nel tempo. Considera cioè il numero di anni, andando all'indietro da quello corrente, in cui si raggiunge il 50% delle citazioni totali ricevute dalla rivista in questione nell'anno precedente. Vi è anche i il *Rate of Cites Index* che rappresenta un indice di qualità del singolo lavoro, nel senso che tanto più il lavoro è citato da altri ricercatori tanto più rilevante viene ritenuto il suo valore scientifico.

Queste misure inevitabilmente risentono delle *mode culturali*, inestirpabili anche negli ambienti scientifici: per esempio sarà più difficile per un ricercatore vedere accettato da una rivista, come si dice, «impattata», un contributo di matrice psicoanalitica in un contesto storico-culturale in cui sia dilagato il paradigma cognitivista stretto; oppure, per essere inclusi in un dibat-

tito corrente, potrà risultare non sufficiente, ma spesso necessario, citare in modo conformista contributi ormai richiamati da ogni ricercatore del settore.

Come vediamo l'oggettività della rilevazione può venire inficiata a monte, nel momento della raccolta dei dati: perché paradossalmete gli autori di contributi veramente innovativi, per definizione, non hanno molti predecessori da citare (in soldoni: non si può essere «freudiani» prima di Freud) ma esplorano, talvolta in modo pionieristico, strade nuove, diverse e perfino ignote nei percorsi di conoscenza.

Le misure basate sugli indici di citazione possono così correre il rischio di risultare, paradossalmente, inversamente proporzionali all'avanzamento scientifico, non riconoscendo adeguatamente e non rilevando tempestivamente il fattore «innovazione».

Ancora, l'impatto scientifico e sociale di una invenzione o scoperta può avere tempi di latenza medio-lunghi e diversificati a seconda del dominio disciplinare: anche cinque o dieci anni prima della sua diffusione su larga scala, come è per la ricerca medica e farmacologica; tempi molto ristretti invece si registrano in genere per l'applicazione delle nano-tecnologie e simili. Queste ultime innovazioni vengono in definitiva rilevate prima dagli indici bibliometrici, in quanto l'avvento di una nuova tecnologia viene testato rapidamente, si diffonde celermente, vista anche la richiesta e la competizione dei mercati internazionali, e di conseguenza ottiene una messe di citazioni maggiori in un lasso di tempo inferiore a quello necessario, per esempio, nella sperimentazione farmacologica o aerospaziale.

Gli indici bibliometrici inoltre, per loro natura, rilevano con più attendibilità la produttività delle scienze «esatte» cosiddette «dure» (neuroscienze, scienze ingegneristiche, mediche, biologiche, fisiche, matematiche, studi nell'ambito della robotica e dell'intelligenza artificiale, ecc.) e non segnalano con la stessa sensibilità la produttività delle scienze umanistiche (compresi interi settori delle scienze umane), in quanto questi settori, in genere, non si appoggiano a livello editoriale su numeri elevati di riviste con *Impact Factor*, eppure hanno le loro sedi editoriali specialistiche (come riviste referate di rilievo, senza o con basso IF), oppure, talvolta, producono più monografie che articoli su rivista.

Come abbiamo visto le tecniche valutative bibliometriche non tengono generalmente conto della differente rappresentatività per area delle riviste censite, rendendo così praticamente impossibile una comparazione tra domini così differenti.

Si viola così una norma statistica basilare, quale la rappresentatività del campione estratto dalla popolazione di riferimento: se i campioni non sono rappresentativi dell'universo da cui vengono estratti non sarà garantita alcuna validità al processo di valutazione. Le tecniche bibliometriche aspirano a

garantire una misurazione oggettiva della produttività scientifica ma cadono nel limite della non confrontabilità dei risultati per settori che dispongono di riviste con IF con distribuzione non omogenea (per esempio, la Psicologia dell'Arte si dispone di pochissime riviste – tra l'altro a basso coefficiente di impatto in quanto sono rivolte ad un numero esiguo di studiosi e quindi destinate comunque ad riscuotere poche citazioni – rispetto in particolare alla Psicologia neurofisiologica, o anche alla Psicologia delle Sviluppo o alla Psicologia sociale), quindi se non si procede ad una preliminare *valutazione pesata del fattore rappresentatività delle riviste con IF negli specifici settori e sotto-settori scientifici disciplinari* non si supera questo consistente limite.

Vogliamo qui solo ricordare e non approfondire, per carità di patria, le condotte, diciamo pure di frode, che l'uso o meglio l'abuso del ricorso all'IF produce nella comunità accademica: dalla notevole mole di auto-citazioni (per le quali si potrebbe allestire un vero e proprio *Self-Citation Index*) al sospetto di un accordo tra *lobby* culturali di citarsi a vicenda, alla mai conteggiabile incidenza delle citazioni negative, volte a dimostrare che il collega citato ha prodotto dati confutabili o si è affannato a dimostrare teorie risultate poi insostenibili.

In sintesi il mito dell'*Impact Factor* si costruisce sull'illusione del criterio dell'oggettività pura nella valutazione e, applicato rigidamente, può produrre vere e proprie distorsioni, nonché aberrazioni come quelle testè verificatesi nell'ambito delle scienze psicologiche che hanno visto praticamente «scomparire» il peso di interi ambiti di ricerca come quella storica, percettologica, fenomenologica nonché, come già ricordato, il settore della psicologia dell'arte e della letteratura: laddove cioè la ricerca non dispone di riviste dotate di alto IF, come succede invece per i settori neuropsicologici.

Occorre sottolineare ancora che la considerazione della specificità dei singoli domini scientifici, ed il rispetto delle differenti tradizioni epistemologiche ed editoriali, garantisce la rappresentatività dei campioni utilizzati, la confrontabilità delle misurazioni e, non ultima, la generalizzabilità dei risultati: di queste cautele metodologiche ogni esercizio di valutazione dei prodotti della ricerca universitaria deve tenere conto per certificare l'affidabilità e la validità delle procedure valutative medesime.

2.2. *Metodi Peer-Review: vantaggi e limiti*

In molti Paesi vengono preferiti per la valutazione della ricerca universitaria i cosiddetti metodi qualitativi tra i quali il *Panel* (di derivazione anglosassone, ora dismesso dalla *Research Assessment Exercis*es o RAE 2008), il *peer-review* (o giudizio dei pari) e l'analisi del beneficio economico (più adatto chiaramente

per le scienze cosiddette «dure» con immediata ricaduta applicativa, piuttosto che per le scienze umanistiche).

I metodi *peer-review* intendono rispondere ad un *criterio qualitativo* nel processo valutativo e si basano sulla valutazione dei prodotti della ricerca individuati da Comitati di Valutatori o dai soggetti stessi che debbono essere valutati. Sono ampiamente trattati in letteratura gli aspetti critici di tale metodologia.

Horrobin ha pubblicato sul *Journal of the American Medical Association* (già nel 1990, n. 263, pp. 1438-1441) un contributo dal titolo «The philosophical basis of peer reviews and the suppression of innovation» («Le basi filosofiche della valutazione *peer-review* e la soppressione dell'innovazione») alludendo al pericolo di conformismo che si può correre se l'originalità non viene rilevata dalla sensibilità del singolo valutatore.

Moxam e Anderson (1992) individuano inoltre importanti limiti dovuti alla *soggettività del valutatore*. Tale valutazione può essere influenzata da conflitti di interesse concreti o ideali, dalla tendenza, anche non consapevole, a valutare comunque meglio i prodotti dei ricercatori più famosi (e qui dobbiamo sottolineare la necessità di maggior rigore metodologico: la valutazione deve implicare la doppia cecità tra valutatore e valutato, oppure la reciproca conoscenza. Il sistema misto è impuro e dà adito all'azione di variabili interferenti come l'effetto dovuto alla fama o suggestione da prestigio); o anche alla forte specializzazione dei contributi da valutare, che necessiterebbero talvolta di competenze altamente settoriali da parte del valutatore.

La precedente valutazione condotta in Italia dal CIVR nel triennio 2001-2003 aveva usato il sistema *peer-review* il quale presenta, come noto in letteratura (D'Angelo, Pugini, & Abramo, 2006), diversi svantaggi in termini di costi, tempi di esecuzione, limiti e oggettività delle misurazioni.

Aspetti critici, oltre che nella scelta dei cosiddetti esperti valutatori del settore (i quali possono essere più aperti o avere anche atteggiamenti pregiudiziali rispetto a particolari gruppi di ricerca o specifici fenomeni allo studio), si rinvengono nella richiesta di una *preventiva selezione*, senza indicazione univoca ed esplicita di vincoli e criteri omogenei, ad opera del singolo ricercatore o docente, dei prodotti da inviare per la valutazione.

Il medesimo criterio ad alto coefficiente di soggettività del valutato è stato adottato alla base dell'attuale procedura VQR 2004-2010: viene cioè operata una scelta ad opera dei singoli studiosi, supervisionata dalle strutture di ricerca, con una blanda indicazione di attenersi alle indicazioni delle tipologie di prodotti più quotati nell'ambito del singolo settore disciplinare, e ciò non esime la campionatura dal ricorso a *criteri impliciti, soggettivi* (come scegliere un prodotto considerato di rilievo ma non collocato in sedi editoriali adeguate, ecc.).

Questa campionatura risente con altà probabilità di *criteri eterogenei*, e di nuovo non rappresentativi dell'universo cui si dovrebbe riferire. Per ovviare a ciò si può solo utilizzare (servendosi finalmente delle banche-dati per le loro potenzialità) l'*intero ammontare dei prodotti scientifici del docente o ricercatore, e classificarli secondo indici espliciti e comuni:* riviste internazionali, riviste nazionali, volumi internazionali, volumi nazionali, ecc. All'estero, il *Center for World University Rankings* (CWR) si è basato su *informazioni reperite su banche dati* ed ha prodotto *una classifica non viziata dalle scelte dei soggetti valutati*.

Anche la campionatura casuale ha un buon indice di rappresentatività, non il migliore, ma molto attendibile sui grandi numeri, occorrerebbe poi una *campionatura stratificata* (per esempio per le diverse fasce di docenza e per le diverse tipologie di prodotti) per garantire anche la presa in considerazione di una componente qualitativa.

La selezione dei prodotti affidata ai singoli docenti e/o alle strutture di ricerca è senz'altro l'anello «debole», in quanto favorisce l'uso di criteri impliciti non misurabili.

Appare necessario a questo proposito aumentare il grado di «controllo» della soggettività del valutato, attraverso procedure intersoggettive più univoche. In questo modo la soggettività controllata – e non arbitraria – diventerà una *risorsa* nella valutazione.

3. Il problema degli «indicatori di qualità» e la necessità del controllo della soggettività nelle procedure di valutazione

3.1. *Il ruolo della soggettività nei processi di valutazione*

Il giudizio valutativo non può che avere una componente soggettiva, la quale va agganciata il più possibile a *criteri oggettivi affidabili e rappresentativi* per garantire la correttezza medesima delle procedure di valutazione. Rivestono un buon valore euristico *indici qualitativi come l'originalità e la rilevanza*, per i quali è però necessario utilizzare scale di misura condivise; sarebbe inoltre utile introdurre nel processo di valutazione delle misure di *follow-up*.

La CRUI ha molto lavorato in passato sulla messa a punto di criteri di efficienza fino a produrne una enorme quantità, talvolta disorientante. Il primo di questi indicatori (detto R1) presenta una validità generale trasversale ai singoli domini e recita così: il grado di efficienza della struttura di ricerca è

dato dal «rapporto tra il numero totale di prodotti della ricerca normalizzati e pesati e il numero pesato di addetti alla ricerca» (espresso dalla formula P/N).

Di nuovo emerge come a livello metodologico non abbia senso confrontare dati grezzi provenienti da popolazioni diverse, ma occorre elaborare (ossia «pesare») in proporzione rispetto all'intero (o universo di riferimento), sia *la quantità e la qualità dei prodotti raccolti in riferimento ad un'area disciplinare, sia la quantità e qualità di «forza lavoro» relativa*.

Ciò è stato fatto nel corso della procedura VQR raggruppando le strutture in «Grandi», «Medie» e «Piccole» ed operando giustamente confronti all'interno delle tre classi; ma non per gli indicatori definiti «dentro» un'area scientifica: essi devono essere integrati per ottenere l'*indicatore finale delle strutture che operano in una pluralità di aree*. L'integrazione richiede di «pesare» gli indicatori di area con un «Peso di area» in modo da rendere l'attribuzione più omogenea. L'ANVUR ha calcolato gli indicatori di area e si è limitata a proporre nella relazione finale un elenco di possibili pesi di area, a partire da quello più diretto che si basa sulla numerosità dei prodotti consegnati. Il risultato dell'integrazione relativa dei Pesi delle diverse aree di una struttura viene demandata al Ministero: ciò inciderà direttamente sulla distribuzione del prossimo FFO.

3.2. *Il controllo della soggettività del valutato nelle procedure per la VQR*

I parametri della VQR prevedevano che ad ogni ricercatore o docente in servizio fosse richiesto di presentare per la valutazione le 3 migliori pubblicazioni del periodo 2004-2010, a proprio giudizio e allineandosi il più possibile – su base volontaria con la raccomandazione di non far fare «brutta figura» all'istituzione di provenienza – ai criteri diffusi nel settore scientifico disciplinare di appartenenza. Non era importante se le 3 pubblicazioni presentate fossero le uniche effettuate o se venissero scelte tra molte; vi era inoltre il divieto di presentare due volte la stessa pubblicazione (nel caso di co-autori) e ciò richiedeva alla struttura attenti confronti per ottimizzare la scelta (a ciascun ricercatore degli enti di ricerca erano richieste 6 pubblicazioni).

La conseguenza registrata è che sono andate meglio le strutture che hanno scelto meglio i «prodotti» da sottoporre alla valutazione, piuttosto che quelle coi prodotti effettivamente migliori: il lavoro di valutazione si è basato quindi non sui dati grezzi ma sulla *rappresentazione mentale* che docenti o ricercatori hanno sviluppato di tali dati. Il lavoro è decisamente apprezzabile e, per chi, come la scrivente, si occupa spesso di contenuti soggettivi per motivi professionali, in qualità di psicologa, appare proprio affascinante: si potrebbe approfondire lo studio vedendo di quanto si discostano le scelte

dei docenti basate sulle loro rappresentazioni mentali dalle indicazioni ministeriali.

Attraverso una breve indagine conoscitiva condotta intervistando colleghi che hanno partecipato alla VQR, sui *motivi* chiamiamoli *personali* che hanno orientato le loro scelte, è stato possibile evidenziare i più frequenti, che riporto solo a titolo di esempio: «[…] ho scelto l'articolo che affronta l'argomento al quale sono più affezionato»; oppure «[…] è un lavoro importante che però ho dato ad un editore locale perché ero sotto concorso e dovevo fare in fretta»; o, ancora, «[…] è un testo che ho presentato ad un Congresso Internazionale ed è stato molto apprezzato, ho ricevuto anche un Premio in quell'occasione, ne sono stato lusingato, anche se, per forza di cose, è pubblicato solo negli Atti di quel Convegno» ed, infine, «[…] è un volume di cinque anni fa ed a quel tempo quella Casa editrice era rinomata, oggi risulta penalizzata perché non ha un grosso sistema di distribuzione a livello nazionale». Altre testimonianze raccolte mettono l'accento su ulteriori aspetti critici: «[…] l'articolo che ho scelto tra i miei migliori è pubblicato in italiano, con il senno di poi, visti i criteri 'a posteriori' utilizzati dall'ANVUR, l'avrei mandato ad una Rivista internazionale in inglese, e avrebbe contato di più» o, ancora, «[…] se avessi saputo che le Curatele non contavano niente, non avrei certo perso tanto tempo negli ultimi dieci anni, ad organizzare Congressi in Italia invitando eminenti esperti stranieri a spese della mia Università, per poi raccogliere in un Volume tutti gli scritti».

Per aumentare il *controllo della soggettività del valutato* sarebbe quindi opportuno basarsi, per i prossimi esercizi di valutazione, su *informazioni reperite su banche dati e selezionate dall'organo preposto alla valutazione nazionale in base a criteri oggettivi ed omogenei*, altrimenti la classifica, come abbiamo detto, sarà viziata dalle personali scelte dei soggetti valutati.

3.3. *Il controllo della soggettività del valutatore nelle procedure VQR*

Come già specificato sono stati coinvolti nella procedura *peer-review* 14.770 revisori, scelti appunto dai GEV (di norma due per prodotto). Essi dovevano esprimere la loro valutazione in base ai criteri resi noti dai GEV, ma non approfonditi previa discussione, prova preliminare, confronto pilota con i colleghi al fine di assicurarsi l'adozione di uno stessa scala di misura per lo stesso criterio. Per non parlare del fatto che l'attribuzione del punteggio andava fatta dopo qualche prova e, meglio, se in un ambiente isolato, non con interferenze imponderabili. I punteggi attribuiti dai revisori, hanno mostrato un certo grado di accordo (è chiaro che un'elevata varianza tra i punteggi attribuiti avrebbe reso praticamente nulla l'intera procedura), ma al fine di

elevare l'affidabilità *dei giudizi* i numerosi *Referees* andavano «addestrati», resi *Referee* «esperti», cioè adeguatamente preparati ad applicare in modo omogeneo i criteri proposti secondo una scala condivisa, questo si auspica venga tenuto in considerazione ai fini del nuovo prossimo esercizio di valutazione della ricerca italiana.

4. Proposte di intervento correttivo per le future procedure di valutazione della ricerca universitaria

Il ricorso al metodo *peer-review* debitamente agganciato ai prodotti preventivamente selezionati (e abbiamo visto che *questa selezione preventiva orienti poi tutta la valutazione*) ed affiancato all'uso di pochi elettivi *indici bibliometrici riconosciuti in campo internazionale* (i quali abbiamo visto non sono omogeneamente riconosciuti e accettati, ma spesso assunti in via transitoria e sottoposti a critiche e revisioni), tende in definitiva a favorire nella distribuzione delle risorse (oggi sempre più esigue) alcuni raggruppamenti disciplinari rispetto ad altri. In tempi di crisi e/o ristrettezze economiche, e quindi di stress, la storia ha registrato spesso un aumento di condotte aggressive e di fenomeni di intolleranza, anche mascherati, dai quali occorre cautelarsi per garantire il progresso scientifico in tutte le sue forme.

Occorre perciò che il calcolo dei vari Indicatori, finora commentati, venga utilizzato con *fine equilibrio*, ricordando sempre che la *valutazione consiste in un'attribuzione di valore* ed, in quanto tale, deve *gestire in modo equilibrato il fattore inalienabile della soggettività*: tenendo cioè conto dei rilievi metodologici sopra ricordati. Solo a queste condizioni l'esercizio di valutazione potrà esprimersi in modo attendibile sulla stima dello sviluppo futuro della ricerca universitaria e sulla necessità di sostenere anche finanziariamente tale sviluppo (sia in termini di FFO, sia di personale ricercatore da acquisire per gli enti di ricerca).

Si impongono a questo punto alcuni interrogativi per un dibattito di approfondimento tra esperti:

1. Quali correttivi metodologici possono migliorare il grado di affidabilità e validità della VQR futura?
2. Occorre garantire un maggior controllo del fattore soggettività nella valutazione della ricerca per ottenere dati più stabili e affidabili. Proposte e riflessioni.
3. Il rigore metodologico, la correttezza e l'appropriatezza delle procedure adottate per sostenere le conclusioni raggiunte caratterizzano la scientifi-

città di un prodotto di ricerca: è quindi un criterio da recuperare nella valutazione della qualità della ricerca universitaria?
4. Come avanzare interpretazioni equilibrate e non ambigue dei dati, non molto stabili, raccolti?
5. Come rendere gli esiti della valutazione utili cioè, dati i punti critici, come riuscire a dare avvio a processi di intervento costruttivo?

Riferimenti bibliografici

ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2013). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale. 30 Giugno 2013*. http://www.anvur.org/rapporto/.

Cuccurullo, F. (2007). *La valutazione della ricerca*. CNEL, http://www.vtr2006.cineca.it.

D'Angelo, C. A., Pugini, F., & Abramo, G. (2006). *La misurazione della produttività scientifica delle università italiane attraverso una metodologia bibliometrica non-parametrica*, Atti del XVII Convegno Nazionale dell'AiLG *Reti, servizi e competitività delle imprese. Sistemi globali e sistemi locali per lo sviluppo*. Roma.

Domenici, G. (2004). La valutazione della ricerca pedagogica. Questioni aperte. *Pedagogia Oggi*, *3*, 16-22.

Horrobin, D. F. (1990). The philosophical basis of peer reviews and the suppression of innovation. *Journal of the American Medical Association*, *263*, 1438-1441.

MacRoberts, H. J., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*, 435-444.

Moxam, H., & Anderson, J. (1992). Peer review. A view from the inside. *Science and Technology Policy*, *5*, 7-15.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometrics methods. *Scientometrics 62*(1), 133-143.