A cura di
Antonio Fasanella
e Fabrizio Martire

Valutazione della ricerca e ricerca sulla valutazione

Riflessioni, analisi e proposte per la VQR

FrancoAngeli
OPEN ACCESS

IL RICCIO E LA VOLPE
Studi, ricerche e percorsi di sociologia

Il riccio e la volpe
Studi, ricerche e percorsi di sociologia

Questa collana ospita, con la più pronunciata apertura tematica e nel pluralismo consapevole delle interpretazioni, indagini empiriche e riflessioni teoriche nell'ambito della sociologia generale.
La sua intestazione richiama un verso di Archiloco che, in uno dei frammenti sopravvissuti, afferma lapidariamente, e in realtà piuttosto oscuramente, che "la volpe sa molte cose, ma il riccio ne sa una grande". Isaiah Berlin, interpretando questa presunta differenza di saperi, scrive, in un saggio degli anni '50, che "esiste un grande divario tra coloro, da una parte, che riferiscono tutto a una visione centrale, a un sistema più o meno coerente e articolato, con regole che li guidano a capire, a pensare e a sentire – un principio ispiratore, unico e universale, il solo che può dare significato a tutto ciò che essi sono e dicono –, e coloro, dall'altra parte, che perseguono molti fini, spesso disgiunti e contraddittori, magari collegati soltanto genericamente, de facto, per qualche ragione psicologica o fisiologica, non unificati da un principio morale ed estetico".
In anni di mutamento sociale e culturale imprevedibilmente accelerato, di "sconfinamenti" e di ibridazioni, questa collana punta dunque a cogliere e documentare le intersezioni e le contrapposizioni, nelle dinamiche sociali, fra l'unitario e il molteplice, il disordinato e il sistemico, il conforme e l'eterogeneo, il caso e la regola: *il riccio e la volpe*, per l'appunto.
Abbandonata la pretesa inattuale di ogni sintesi semplice, difficilmente la sociologia potrebbe oggi sottrarsi a questo lavoro paziente di ricostruzione.
La molteplicità delle tematiche affrontate e la pluralità delle prospettive trovano, peraltro, una precisa composizione unitaria nella ferma e rigorosa opzione disciplinare che ispira la collana stessa, e cioè nella puntigliosa rivendicazione della sociologia come disciplina costantemente attenta all'integrazione tra teoria e ricerca, al rigore logico-metodologico delle procedure, al rispetto della fondamentale esigenza di pubblicità e controllabilità dell'indagine scientifica.
Sulla base di questi convincimenti di natura teorico-metodologica, e nel costante richiamo alla responsabilità sociale di ogni disciplina scientifica, la collana si propone di fornire a studiosi, a studenti e a operatori strumenti qualificati di riflessione e di intervento.

 ISBN 9788835103219

Copyright © 2020 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835103219

A cura di
Antonio Fasanella
e Fabrizio Martire

Valutazione della ricerca e ricerca sulla valutazione

Riflessioni, analisi e proposte per la VQR

FrancoAngeli

 ISBN 9788835103219

Questo volume è stato pubblicato con il contributo del Dipartimento di Comunicazione e Ricerca Sociale di Sapienza Università di Roma.

 ISBN 9788835103219

Indice

Introduzione, di *Antonio Fasanella* e *Fabrizio Martire* pag. 7

1. Esperienze di valutazione della ricerca scientifica, di *Marco Palmieri* » 13
1.1. Introduzione » 13
1.2. Le prime esperienze di valutazione in Italia » 17
1.3. La valutazione triennale della ricerca, la Vtr 2001-2003 » 21
1.4. La valutazione della qualità della ricerca, la Vqr 2004-2010 » 25
1.5. La valutazione della qualità della ricerca, la Vqr 2011-2014 » 28
1.6. L'Anvur » 35
1.7. Una comparazione dei principali sistemi europei di valutazione della qualità della ricerca » 37

2. La definizione del concetto di qualità della ricerca, di *Antonio Fasanella* » 48
2.1. Introduzione » 48
2.2. Determinatezza e uniformità d'uso del concetto di qualità » 49
2.3. Un tentativo di chiarificazione concettuale » 58

3. I Gev e i revisori, di *Lorenzo Barbanera* » 65
3.1. I Gruppi di Esperti Valutatori » 69

 ISBN 9788835103219

3.1.1. La composizione dei gruppi pag. 69
3.1.2. L'importanza dell'eterogeneità dei gruppi » 71
3.1.3. L'assegnazione dei prodotti agli Ev » 74
3.1.4. Strategie di *accountability* » 76
3.2. I revisori » 79
3.2.1. Il processo di selezione » 79
3.2.2. Il *matching* prodotto-revisore » 83
3.2.3. L'*accountability* recitativa » 85
3.3. Conclusioni » 91

4. La scheda di valutazione dei prodotti scientifici, di *Federica Floridi* » 93
4.1. Introduzione » 93
4.2. La struttura della scheda Vqr 2011-2014 » 94
4.3. Gli oggetti multipli e le conseguenze della sotto-determinazione semantica » 98
4.4. La manipolazione *ex post* delle classi di merito: una questione di qualità » 102
4.5. I giudizi obbligatori: un'occasione mancata » 108

5. Formulazione e sintesi dei giudizi, di *Federica Fusillo* » 112
5.1. La classificazione della qualità della ricerca » 112
5.2. La sintesi del giudizio del singolo revisore » 117
5.3. L'attribuzione della classe di merito finale » 129
5.4. Conclusioni » 137

6. Fare ricerca sulla Vqr, di *Fabrizio Martire* » 139

Riferimenti bibliografici » 149

Curatori e autori » 157

 ISBN 9788835103219

Introduzione

di Antonio Fasanella *e* Fabrizio Martire

Nell'introdurre il presente volume è doveroso nei confronti del lettore delineare le tempistiche del corso di vita dell'oggetto di riflessione a cui ruota intorno: la procedura di Valutazione della Qualità della Ricerca.

La Vqr che viene presa in analisi è quella relativa alla valutazione dei prodotti della ricerca scientifica pubblicati tra il 2011 e il 2014, procedura che – al momento dell'uscita del presente volume – è conclusa da tre anni con la pubblicazione del rapporto finale nel febbraio 2017. Con la recente emanazione del d.m. n. 1110 del 2019, recante le "Linee guida per la valutazione della qualità della ricerca (Vqr) 2015-2019", e il successivo bando ha preso ufficialmente avvio il nuovo esercizio valutativo in materia. Tuttavia, le procedure che si intendono adottare per l'ultimo quinquennio non sembrano presentare sostanziali cambiamenti, tali da invalidare le considerazioni esposte nel testo, il quale mira a dare voce ad una prospettiva che vuole contribuire a un dibattito costruttivo e volto al miglioramento. Le critiche, le proposte e i contenuti del presente volume possono essere considerati quindi ancora utili per una lettura analitica della più generale procedura di Valutazione della qualità della ricerca.

Ognuno dei capitoli che compone il volume affronta un aspetto specifico della procedura analizzata, con l'intento di restituire al lettore un quadro il più completo possibile dell'esercizio di valutazione in esame.

Nel primo capitolo l'autore ricostruisce il quadro normativo che introduce nel sistema universitario italiano la valutazione della qualità

 ISBN 9788835103219

della ricerca, e in particolare dei prodotti di ricerca pubblicati da docenti e ricercatori universitari. Gli interventi legislativi che si sono succeduti nel tempo hanno un denominatore comune: legare parte dell'ammontare dei fondi pubblici destinati all'università all'esito degli esercizi di valutazione. Negli anni si assiste a un costante incremento della cosiddetta quota premiale del Fondo di Finanziamento Ordinario, che muta il rapporto tra lo Stato e il sistema universitario: lo Stato cessa di regolare a monte l'entità dei fondi da assegnare a ciascuna università, per diventare un valutatore che a valle attribuisce le risorse premiali in base ai risultati conseguiti degli atenei in occasione degli esercizi di valutazione. Nel corso del capitolo sono passate in rassegna alcune esperienze di valutazione della qualità della ricerca del sistema universitario italiano. In particolare, si dà spazio alla Vtr 2001-2003 (Valutazione triennale della ricerca, 2001-2003), alla Vqr 2004-2010 (Valutazione della qualità della ricerca, 2004-2010) e alla Vqr 2011-2014, che è per l'appunto oggetto della trattazione di questo volume. Si traccia così un percorso che a partire dagli anni Novanta, con l'istituzione dell'Osservatorio per la valutazione del sistema universitario, poi sostituito dal Comitato nazionale per la valutazione del sistema universitario e dal Comitato di indirizzo per la valutazione della ricerca, giunge ai nostri giorni, in cui vede la luce l'Anvur, Agenzia nazionale di valutazione del sistema universitario e della ricerca. Nell'ultima parte del capitolo si presentano i risultati di una comparazione internazionale dei principali sistemi di valutazione della qualità della ricerca, dalla quale emergono differenze sostanziali negli approcci adottati dai maggiori paesi europei.

Il capitolo 2 affronta il tema della definizione della nozione di qualità della ricerca, resa nei documenti costitutivi della Vqr. L'analisi svolta evidenzia una certa vaghezza e conseguente ambiguità del concetto adottato ai fini della valutazione, che si traducono nella difficoltà di escludere dalla definizione, e perciò dall'uso del concetto stesso, elementi esterni, potenzialmente spuri, estranei ad esso, riconducibili alla soggettività mutevole del singolo valutatore. In tal modo, si assiste alla violazione del principio di determinatezza e uniformità d'uso, risultando così tendenzialmente pregiudicata la possibilità stessa di una comparazione tra i giudizi di due diversi valutatori espressi nello stesso momento, ovvero tra i giudizi di uno stesso valutatore espressi in momenti diversi. Nel corso della trattazione, viene formulata una

 ISBN 9788835103219

proposta di chiarificazione concettuale della nozione in esame, mediante l'individuazione e l'analisi lessicale di alcune dimensioni riconducibili ad essa (*originalità*; *accuratezza metodologica*; *appropriatezza della scrittura*; *riferimento alle fonti*; *adeguatezza della trattazione*; *reputazione e selettività della rivista*). Il capitolo si chiude con un'avvertenza circa la necessità di considerare il percorso che conduce alla chiarificazione e alla condivisione del concetto a fini valutativi come massimante inclusivo e negoziale, in grado di coinvolgere tutti gli attori interessati alla valutazione, autori, revisori, esperti valutatori, decisori. Ciò proprio al fine di ripristinare il principio di determinatezza e uniformità d'uso e la legittimità stessa dei processi valutativi.

Il capitolo 3 costituisce un'analisi critica delle procedure utilizzate per la composizione dei gruppi Gev e per la gestione dei prodotti durante tutto il processo valutativo; dallo smistamento interno dei contributi alla selezione dei *referees*, fino alla risoluzione delle controversie. Nello specifico, si illustrano anzitutto le problematiche emerse in merito al reclutamento degli esperti valutatori per l'Area 14, evidenziando da un lato, gli aspetti negativi legati alla scarsa eterogeneità del gruppo in termini di competenze disciplinari, dall'altro, il sovraccarico di lavoro dovuto alla necessità di amministrare un'ingente quantità di prodotti. In seguito, l'attenzione si focalizza sulle strategie di *matching* tra prodotto e revisore, che troppo spesso hanno messo i *referees* nella condizione di dover valutare i contributi senza possedere le conoscenze necessarie. Di conseguenza, seguendo un principio orientato a valorizzare le abilità e le competenze di tutti gli attori coinvolti nella Vqr, si avanzano alcune proposte per arginare i *vulnera* evidenziati, partendo in primo luogo dalla revisione delle linee guida, principale strumento di condivisione delle pratiche la cui realizzazione non sembra ormai poter prescindere dal contributo attivo della comunità scientifica. Infine, si esprime la necessità di acuire gli sforzi al fine di garantire un adeguato livello di *accountability*, non solo per ragioni di trasparenza, ma anche per favorire il costante miglioramento della Vqr in tutte le sue fasi.

Il capitolo 4 è incentrato sulla scheda impiegata dai *referees* dell'Area 14 per la valutazione dei prodotti scientifici sottoposti a Vqr 2011-2014. Dopo aver analizzato la struttura della scheda, il contenuto delle domande che la compongono e le caratteristiche della tecnica di *sca-*

 ISBN 9788835103219

ling impiegata, vengono presentate le conseguenze della sotto-determinazione semantica e della presenza di oggetti multipli nel testo delle domande. Successivamente, vengono trattate la questione della possibilità di modificare *ex post* l'attribuzione del prodotto a una classe di merito e la questione della gestione della sezione dedicata all'espressione obbligatoria di un giudizio esteso che avrebbe dovuto ricoprire un ruolo strategico nel più generale processo di valutazione, anche rispetto ad alcune fasi maggiormente problematiche come, ad esempio, la possibilità di cambiare i punteggi. Il percorso proposto nel capitolo 4 è teso a dimostrare come, da un punto di vista prettamente metodologico, l'assenza di operazioni di chiarificazione concettuale e di specificazione di significato del concetto di qualità della ricerca, così come emerso nel capitolo 2, possano aver generato difficoltà e, in alcuni casi, prodotto distorsioni nella fase di rilevazione empirica del concetto.

Nel capitolo 5 viene affrontata la procedura di attribuzione di una classe di merito ai prodotti sottoposti a valutazione, a partire dai giudizi forniti dai *referees*. L'analisi del sistema di classificazione della qualità adottato nella Vqr è stata portata avanti con lo scopo di evidenziare le conseguenze, positive e/o negative, dell'adozione di una strategia metodologica piuttosto che di un'altra. Il primo passo è stato analizzare la definizione semantica formulata per le classi di merito (*eccellente, elevato, discreto, accettabile, limitato*) costruita sulla base dei tre criteri di qualità della ricerca, mettendo in evidenza come in realtà esse restituiscano dei tipi ideali, ossia situazioni in cui le valutazioni fornite sui tre criteri risultino perfettamente concordanti, e vengano tralasciati i casi di valutazioni discordanti di una o più classi. Successivamente, si è entrati nel merito dell'attribuzione di una classe di merito da parte di un *referee* e della sintesi del giudizio finale sul prodotto valutato. Si è cercato di mettere in luce come la poca trasparenza in questa delicata fase abbia reso poco chiari quali siano stati i criteri procedurali di sintesi dei giudizi e, conseguentemente, di gestione dei casi discordanti di una o più classi, che hanno rappresentato la maggioranza delle valutazioni gestite dai Gev. Filo conduttore di questa riflessione sono stati i presupposti metodologici per l'adozione di una procedura di sintesi numerica oppure tipologica, senza tralasciare l'importanza sostanziale e pratica che riveste una buona fase di negoziazione dei significati e di condivisione dei codici interpretativi dei concetti analizzati e delle definizioni utilizzate.

A partire dalle problematiche sollevate dai diversi autori nel corso del volume, nel sesto capitolo si dà spazio ad alcune idee di ricerca valutativa sulla Vqr. È nostra intenzione proporre l'avvio di un programma di ricerca che suggerisca soluzioni alle difficoltà emerse dalla nostra analisi sull'impianto dell'ultima Vqr. Nel capitolo si sollevano diverse domande di ricerca possibili. In primo luogo, va compreso se la scelta del Miur di indirizzare i flussi di finanziamento verso le università che ottengono i risultati migliori nella Vqr stia incentivando gli atenei meno produttivi a incrementare la propria *performance*, o se al contrario stiamo assistendo un acuirsi delle distanze tra i primi e chi invece fa fatica a far propri i nuovi criteri di valutazione imposti dall'Anvur. Il capitolo illustra anche idee di ricerca sulla selezione dei Gev e dei revisori dei prodotti: quale dovrebbe essere la numerosità di un Gev tale da garantire la giusta rappresentatività di tutti i settori scientifico-disciplinari di ciascun'area scientifica? Quali criteri dovrebbero essere adottati dai Gev nella scelta dei revisori più appropriati per ciascun prodotto inviato a valutazione? Un'altra questione cruciale da indagare riguarda i criteri adottati *di fatto* dai revisori nella valutazione dei prodotti, in relazione a quelli indicati nella scheda di valutazione messa a loro disposizione. Molti altri interrogativi di ricerca sono presentati nel capitolo; lo scopo è stimolare la comunità scientifica, e gli attori istituzionali coinvolti, ad avviare un programma di ricerca e un processo di apprendimento condiviso in vista dei futuri esercizi di valutazione della qualità della ricerca.

 ISBN 9788835103219

 ISBN 9788835103219

1. Esperienze di valutazione della ricerca scientifica

di Marco Palmieri

1.1. Introduzione

In Italia l'esigenza di valutare le università comincia ad essere avvertita alla fine degli anni Ottanta a seguito del riconoscimento della loro autonomia nei confronti degli organi centrali di governo. La legge n. 168 del 1989, che ha sancito l'autonomia didattica, scientifica, organizzativa delle università italiane, già prevista dalla Costituzione, ha messo in moto un processo di riforma del sistema universitario proseguito nei decenni successivi.

Punto di svolta è la legge n. 180 del 2008 in cui si legge che, «a decorrere dall'anno 2009, al fine di promuovere e sostenere l'incremento qualitativo delle attività delle università statali e di migliorare l'efficacia e l'efficienza nell'utilizzo delle risorse, (…) una quota non inferiore al 7 per cento del fondo di finanziamento ordinario con progressivi incrementi negli anni successivi, è ripartita prendendo in considerazione la qualità dell'offerta formativa e i risultati dei processi formativi e la qualità della ricerca scientifica». La legge n. 69 del 2013 è poi intervenuta sulle modalità di attribuzione della quota premiale del fondo di finanziamento ordinario all'università negli anni a venire.

Questo intervento legislativo cambia nel profondo il meccanismo di finanziamento statale: parte dell'ammontare dei fondi pubblici destinati all'università viene legato alle prestazioni rese e al valore prodotto dai singoli atenei, che gestiscono i finanziamenti ricevuti in piena autonomia. Nel nuovo sistema di redistribuzione dei fondi pubblici diventa centrale la dimensione della *performance* degli atenei da sottoporre a valutazione (Reale, 2013a).

 ISBN 9788835103219

La didattica, più in generale la formazione, e la ricerca sono tra gli aspetti che entrano nel processo di valutazione. Dall'analisi dei decreti ministeriali del Ministero dell'Istruzione, dell'Università e della Ricerca (Miur) emerge che: la quota premiale del fondo di finanziamento ordinario (Ffo) cresce costantemente dal 2009 al 2019; una parte sempre più consistente del Ffo è redistribuito tenendo conto degli esiti degli esercizi di valutazione della qualità della ricerca prodotta dal sistema universitario nazionale[1].

Nel 2009 la quota premiale è stabilità al 7% del Ffo, il 50% della quale è assegnata secondo i risultati della Vtr 2001-2003[2]. Nel 2010 e nel 2011 la quota premiale è pari al 12% del Ffo, e nel 2012 tale quota sale al 13%; in generale nel triennio il 12% del fondo premiale è redistribuito tenendo conto della Vtr 2001-2003. Nel 2013 la quota premiale è al 13,5% del Ffo, il 66% della quale è assegnata tenendo conto degli esiti della Vqr 2004-2010. Nel 2014 il fondo premiale sale al 18% del Ffo, il 90% del quale è ripartito in base agli esiti della Vqr 2004-2010. La quota premiale per l'anno 2015, e anche per il 2016, è il 20% del Ffo; l'85% di tale importo è suddiviso secondo gli esiti della Vqr 2011-2014. Nel 2017 il Miur ha destinato ben il 22% del Ffo al fondo premiale, e l'85% di questo è stato erogato in funzione dell'ultima Vqr. La quota premiale sale al 24% nel 2018 e al 26% nel 2019; il 60% di questa è redistribuita secondo i risultati della Vqr 2011-2014.

In questo mutamento di prospettiva ha giocato un ruolo importante la diffusione delle idee del *new public management*, una corrente di pensiero che ha orientato le scelte del governo nella direzione di importare nel settore pubblico metodi di *management* aziendale (Reale e Pennisi, 2012). Secondo questo approccio, la valutazione dei risultati è uno strumento di regolazione del funzionamento dei sistemi pubblici, perché supporta un programma di incentivazione che ha il fine di orientare i comportamenti dei soggetti valutati. Le università, in

[1] Vista la delicatezza dell'argomento, chi scrive consiglia l'avvio di un programma di ricerca valutativa che analizzi l'impatto della redistribuzione premiale dei fondi sull'operato degli atenei italiani. Al riguardo vedi il capitolo 6.

[2] La Vtr 2001-2003 è stata la prima esperienza di valutazione della qualità della ricerca nelle università italiane (sulla Vtr 2001-2003 vedi il par. 1.3), poi seguita da altri due esercizi di valutazione, la Vqr 2004-2010 (sulla Vqr 2004-2010 vedi il par. 1.4) e la Vqr 2011-2014, che è per l'appunto oggetto di questo libro.

quanto istituzioni del settore pubblico, sono quindi chiamate a rendicontare le loro attività in vista della ripartizione degli stanziamenti statali.

Non solo. La legge n. 537 del 1993 istituisce l'autonomia finanziaria degli atenei, grazie alla quale cadono i vincoli di destinazione contabile delle università nei confronti dello Stato finanziatore: i fondi pubblici non sono più definiti dal Ministero in voci di spesa univoche, che ne limitano la destinazione d'uso, ma entrano a far parte di un unico *budget*, il fondo di finanziamento ordinario; ciascun ateneo gestisce autonomamente la quota parte che gli viene riconosciuta.

L'autonomia di cui godono le università le obbliga all'esercizio di responsabilità nella gestione delle risorse pubbliche assegnate. Ciò implica la necessità di rendere conto pubblicamente del loro operato e dei risultati conseguiti per favorire il controllo democratico sulle modalità di gestione della cosa pubblica (*accountability*). In questo senso, la valutazione è strettamente connessa alle responsabilità di cui sono investite le università in seguito al riconoscimento della loro autonomia finanziaria. Valutazione come controllo di affidabilità e come mezzo per favorire un esercizio responsabile dell'autonomia.

> La valutazione si afferma come corollario indispensabile del riconoscimento alle università di autonomia di tipo sostanziale e non meramente procedurale. Essa deve, infatti, nelle intenzioni del legislatore, consentire alle università di acquisire consapevolezza del proprio ruolo e delle proprie funzioni, favorire l'auto-riflessione, contribuire a rendere evidenti le criticità che necessitano d'intervento, e consentire di conoscere i risultati prodotti e il loro utilizzo, e chi concorre alla loro produzione (Reale, 2013b, p. 148-149).

Si assiste al passaggio da una forma di regolamentazione tradizionale, in cui lo Stato stabilisce *ex ante* l'allocazione delle risorse ed è ultimamente responsabile della loro gestione, a una in cui la responsabilità è trasferita agli atenei, che devono però dimostrare di saper gestire l'autonomia sottoponendosi a una valutazione *ex post* dei loro risultati. Lo Stato abbandona progressivamente la figura di finanziatore che controlla il sistema agendo a monte sulle quote di risorse assegnate ai diversi atenei, per assumere progressivamente il ruolo di valutatore che a valle attribuisce le risorse premiali in base al buono o cattivo uso che fin lì ne è stato fatto.

Da questa prospettiva, la valutazione delle università serve a dare

 ISBN 9788835103219

prova della loro affidabilità a tutti coloro che, a vario titolo, sono portatori di interessi nei confronti delle istituzioni accademiche: non solo gli organi centrali di governo ma anche le imprese del territorio e, più in generale, la cittadinanza (Ribolzi, 2013).

Il modello di *accountability* configuratosi nel contesto italiano ha sollevato però le resistenze di molti docenti e ricercatori, per via di un sistema di valutazione che spingerebbe a conformare le attività di ricerca e didattica delle università a criteri di qualità imposti dall'esterno.

> In termini più generali, usi della valutazione di tal fatta comportano la sostituzione del sistema di regolazione specifico al campo della scienza con un nuovo sistema di normazione eteronoma: il potere economico e politico verrebbe così a sottrarre spazio ad altre forme di potere più consolidate nell'accademia, quali il potere istituzionale e istituzionalizzato (occupazione di posti di potere) e quello specifico della scienza (fondato sul prestigio e l'autorevolezza). La qual cosa è connessa a una perdita di autonomia del campo scientifico (Chessa e Vargiu, 2011, p. 23).

La valutazione è stata così percepita dal mondo accademico non come una procedura per incrementare la responsabilità istituzionale dall'università italiana, ma come uno strumento per controllare la conformità degli atenei agli standard prestazionali richiesti dalle autorità governative (Palumbo e Pennisi, 2011). Parte della comunità accademica è dell'opinione che in questo modello di valutazione l'innalzamento della qualità delle università è una finalità perseguita solo indirettamente, tramite un sistema di allocazione delle risorse che mira a incoraggiare i comportamenti conformi rispetto agli standard. Ciò ha favorito condotte di tipo opportunistico che puntano a garantire il rispetto formale, e non sostanziale, dei requisiti richiesti. Se però la conformità innescata è solo formale, la valutazione perde la sua più importante funzione, ossia favorire il miglioramento di didattica e ricerca mediante un processo di apprendimento (Rebora, 2013).

> Quando scopo della valutazione è controllare il settore o verificarne la conformità alle politiche del governo o assicurare la rendicontazione (di solito in riferimento ai contribuenti), e la metodologia comprende qualche forma di ispezione, sebbene nella forma apparentemente benevola della *peer review*, allora non c'è alcun segnale di democrazia. Tutto il sistema appare autocratico e all'interno delle università è percepito come autocratico (Harvey, 2008, p. 9).

 ISBN 9788835103219

Secondo Valentini (2013), piuttosto che un'attività burocratica di rendicontazione a un organo di controllo esterno, la valutazione può costituire un'insostituibile occasione di riflessione, un momento di condivisione dei risultati ottenuti che può favorire un processo di apprendimento da parte della comunità accademica. Da questa prospettiva gli esiti del processo valutativo rappresentano non tanto un supporto al sistema premiale di allocazione delle risorse quanto un bacino informativo cui gli atenei possono attingere per definire le loro politiche nei campi della didattica e della ricerca. La logica del miglioramento porta cioè ad attribuire alla valutazione la funzione di rendere espliciti operato e risultati dei singoli atenei per promuovere un confronto dialettico tra le diverse esperienze del mondo accademico.

Cipriani (2013) osserva che un sistema orientato al miglioramento della qualità degli atenei più che alla ricerca di prove della loro affidabilità favorisce un meccanismo di autoregolazione da parte della comunità accademica. Ciò può accadere se la valutazione non è percepita dagli attori come attività inquisitoria cui reagire difensivamente, ma come pratica riflessiva che spinge alla ricerca di un cambiamento. La collaborazione della comunità accademica alla costruzione dell'edificio valutativo diventa così la chiave per la diffusione di una cultura della valutazione capace di favorire un processo di responsabilizzazione di tutti gli attori coinvolti.

Per costruire una simile rappresentazione della valutazione è opportuno coinvolgere docenti e ricercatori nel processo di definizione dei criteri con cui valutare la loro attività, riflettendo sulle dimensioni della qualità della ricerca per esplicitare, formalizzare e proporre pratiche di valutazione adeguate ai diversi contesti scientifico-disciplinari.

Proprio questo è lo spirito che ha guidato le riflessioni raccolte in questo volume, dedicato all'ultimo esercizio di valutazione della qualità della ricerca, la Vqr 2011-2014.

1.2. Le prime esperienze di valutazione in Italia

La valutazione entra nel mondo accademico italiano negli anni Novanta, in seguito all'approvazione della legge n. 537 del 1993, successivamente riformata dalla legge n. 370 del 1999. Le prime novità importanti riguardano l'istituzione dei Nuclei di valutazione interna e

dell'Osservatorio per la valutazione del sistema universitario. I primi sono organismi interni agli atenei «con il compito di verificare, mediante analisi comparativa dei costi e dei rendimenti, la corretta gestione delle risorse pubbliche, la produttività della ricerca e della didattica, nonché l'imparzialità ed il buon andamento dell'azione amministrativa […] La relazione annuale dei Nuclei di valutazione interna è trasmessa al Ministero dell'Università e della ricerca scientifica e tecnologica anche ai fini della successiva assegnazione delle risorse» (legge n. 537 del 1993).

La legge affida ai Nuclei il ruolo delicato di valutatori interni all'università, con il compito di valutare la produttività dell'attività degli atenei. Il Nucleo è lo strumento di valutazione del sistema universitario italiano, che verifica la coerenza tra gli obbiettivi programmati, le risorse impiegate, e i risultati raggiunti da ciascun ateneo. Al Nucleo spetta dunque il compito oneroso di valutare l'intera gestione dell'ateneo (didattica, amministrativa, di ricerca), ivi compresa l'analisi e la valutazione della direzione politica degli atenei.

Questa innovazione legislativa fatica però ad andare a sistema. In un documento del 1997 che riporta le conclusioni di un lavoro di monitoraggio dell'attività dei Nuclei a quattro anni dalla loro istituzione (Osservatorio, 1997, Doc 5/97) si legge che: in diversi atenei la nomina dei membri dei Nuclei di valutazione interna risponde solo a un obbligo formale, poiché i Nuclei non redigono la relazione annuale prevista dalla legge; la durata dell'incarico dei membri, e il relativo compenso, non sono adeguati all'impegno richiesto per svolgere un compito così oneroso; i Nuclei sono costituiti esclusivamente da membri interni all'ateneo, scelti dal rettore tra i docenti interni (a scapito di esperti esterni) senza che la loro autonomia sia adeguatamente tutelata.

L'altra importante novità della legge n. 537 del 1993 è la nascita di un Osservatorio permanente, istituito poi nel 1996. L'Osservatorio è un organo tecnico con il compito «di valutare i risultati relativi all'efficienza e alla produttività delle attività di ricerca e di formazione; di valutare i piani di sviluppo e riequilibrio del sistema universitario». Il suo compito principale è redigere una relazione annuale sullo stato della valutazione del sistema universitario italiano, elaborata sulla base delle relazioni preparate annualmente dai Nuclei di valutazione interna.

 ISBN 9788835103219

> Con l'Osservatorio si passa da un sistema di valutazione costruito come sommatoria di soggetti decentrati (i Nuclei), più o meno omologati e tenuti in rete da basi di dati confrontabili, a un sistema duale, che combina l'attività di valutazione delle singole università con quella del sistema universitario nel suo complesso (Rizzi e Silvestri, 2002, p. 6).

Il neonato Osservatorio non sostituisce né limita il lavoro dei Nuclei di valutazione, anzi ne rafforza l'autonomia e l'indipendenza dagli atenei, essendo un organo tecnico composto da studiosi autorevoli, che raccoglie ed elabora il lavoro dei singoli Nuclei di valutazione e relaziona al Miur sullo stato della valutazione dell'università italiana. Si assiste così alla nascita del primo organismo nazionale di valutazione del sistema universitario, cui sono assegnati anche numerosi altri compiti: la valutazione dei piani triennali di sviluppo, i pareri sulle nuove università e sui nuovi dottorati di ricerca, la valutazione sulla riforma dell'autonomia didattica, ecc. Inoltre, l'Osservatorio suggerisce al Ministero come ripartire, o meglio riequilibrare, le quote del Ffo da assegnare agli atenei secondo il criterio delle pari opportunità da garantire a tutti gli atenei, con particolare attenzione a quelli che operano in contesti con condizioni socioeconomiche svantaggiate (Matarazzo, 2018).

Il ruolo e i compiti dei Nuclei di valutazione interna e dell'organismo centrale di valutazione del sistema universitario sono stati poi ridefiniti dalla legge n. 370 del 1999 che, facendo tesoro dell'esperienza di sei anni, cerca di valorizzare i punti di forza e correggere le inevitabili criticità. In primo luogo, vengono confermati i compiti già assegnati ai Nuclei di valutazione dalla precedente riforma (redazione della relazione annuale di accompagnamento al consuntivo di ateneo e valutazione di congruità tra gli obiettivi dichiarati e i mezzi indicati nell'ambito della programmazione triennale), conferendo loro anche altre funzioni che prima erano di competenza dell'Osservatorio (rilevazione delle opinioni degli studenti e redazione del rapporto sulla valutazione della didattica, valutazione dei requisiti per l'attivazione dei corsi di dottorato, valutazione della riforma degli ordinamenti didattici). Inoltre, viene rafforzata l'autonomia dei Nuclei rispetto agli atenei che sono chiamati a valutare, i quali da quel momento in poi sono obbligati a fornire loro le informazioni richieste per la valutazione dell'ateneo. Sono previste anche sanzioni (nel meccanismo di attribuzione dei fondi) per le università che non mettono il proprio Nucleo di

valutazione interna in condizione di lavorare con professionalità e imparzialità.

Rimane però un limite sostanziale all'azione del Nucleo: l'impossibilità di imporre all'ateneo modifiche ai progetti sulla base dei risultati dell'attività valutativa; solitamente il Nucleo non viene coinvolto nemmeno nella stesura iniziale, neanche per dare valutazioni informali. Al contrario, far partecipare il Nucleo alla scrittura del progetto sarebbe il «modo di procedere più vantaggioso sia per l'ateneo, che aumenterebbe le probabilità di successo dei propri progetti poiché in essi sarebbero già contenute le indicazioni del Nucleo interno, sia per il Nucleo che riuscirebbe a svolgere un positivo ruolo di valutazione senza scontrarsi con gli organi dell'ateneo, sia per il Comitato e il Ministero che riceverebbero progetti migliori» (Rizzi e Silvestri, 2002, p. 14).

La legge n. 370 del 1999 prevede la sostituzione dell'Osservatorio per la valutazione del sistema universitario con il Comitato nazionale per la valutazione del sistema universitario (Cnvsu), cui sono affidati i compiti fino a quel momento svolti dall'Osservatorio. La novità principale sta nel rapporto più stretto che il Comitato è tenuto a instaurare con i Nuclei di valutazione, fissandone gli obiettivi di lavoro e le informazioni che devono comunicare annualmente al Comitato, e preparando un programma annuale di valutazione esterna delle università. Si assiste, dunque, al rafforzamento dell'architettura piramidale dell'intero sistema di valutazione, al cui apice c'è il Cnvsu, che dialoga direttamente con il Ministero e che è il punto di riferimento di tutti i Nuclei di valutazione interna. Inoltre, questo suo ruolo è rafforzato dall'assenza di un legame orizzontale, a rete, tra i Nuclei medesimi (Scarpitti, 2001).

Dal punto di vista legislativo, lo sviluppo del sistema di valutazione dell'università italiana corre parallelo a quello dei sistemi di controllo interni alla pubblica amministrazione, istituiti dalla legge Bassanini n. 59 del 1997 e relativi decreti attuativi, in particolare il decreto n. 286 del 1999 di «riordino e potenziamento dei meccanismi e strumenti di monitoraggio e valutazione dei costi, dei rendimenti e dei risultati dell'attività svolta dalle amministrazioni pubbliche». In tale legge, e in quelle che successivamente la modificano parzialmente, si istituisce la chiara distinzione tra controllo di gestione, demandato a un organo interno all'amministrazione per controllare l'economicità, efficienza ed efficacia della gestione (nel sistema universitario tale ruolo è svolto dai Nuclei di

 ISBN 9788835103219

valutazione interna), e controllo sulla gestione, affidato a un organo esterno come la Corte dei Conti, cui è demandata «l'analisi, preventiva e successiva, delle conseguenze e/o degli eventuali scostamenti tra le missioni affidate dalle norme, gli obiettivi operativi prescelti, le scelte operative effettuate e le risorse umane, finanziarie e materiali assegnate, nonché l'identificazione degli eventuali fattori ostativi, delle eventuali responsabilità per la mancata o parziale attuazione, dei possibili rimedi».

1.3. La valutazione triennale della ricerca, la Vtr 2001-2003

La Vtr 2001-2003 è la prima vera esperienza di valutazione della qualità della ricerca prodotta dal sistema universitario italiano, approvata dal Miur, che con il d.m. n. 2206 del 2003 ha affidato al Comitato di indirizzo per la valutazione della ricerca (Civr) l'attuazione e l'organizzazione dell'intero esercizio di valutazione. Il Civr è istituito dal Miur con d.m. n. 204 del 1998, in cui si legge che il Comitato «è composto da non più di 7 membri, anche stranieri, di comprovata qualificazione ed esperienza, scelti in una pluralità di ambiti metodologici e disciplinari. Il comitato opera 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».

Diversi sono i compiti assegnati al Civr in occasione dell'organizzazione della Vtr: a) la scrittura delle linee guida; b) la definizione dei criteri per la composizione dei panel di area e la proposta al Miur dei componenti da nominare; c) la valutazione dei dati trasmessi dalle strutture; d) la promozione di incontri periodici con i componenti dei panel; e) la valutazione delle relazioni e dei rapporti inviati dai panel; f) l'organizzazione di eventuali audizioni con le parti interessate; g) la relazione finale per singola struttura; h) relazioni periodiche di valutazione del Sistema Nazionale della Ricerca; i) la pianificazione dell'utilizzazione delle risorse ad esso destinate.

Alla Vtr sono invitati a partecipare le università e gli enti di ricerca pubblici e anche privati convenzionati con il Ministero.

Due sono le principali linee guida decise dal Civr per la Vtr 2001-2003: la volontarietà delle strutture di ricerca di aderire al programma di

valutazione[3], l'attenzione alla produzione scientifica di alta qualità pubblicata dal 2001 al 2003. Riguardo a quest'ultimo aspetto, il Civr ha deciso di concentrare la valutazione su pochi prodotti considerati l'eccellenza scientifica. A ciascun ateneo, infatti, è stato chiesto di selezionare un numero esiguo di prodotti, pari al 50% del numero medio nel triennio 2001-2003 dei docenti e ricercatori in esso operanti. Per le strutture non vi è stato alcun obbligo di presentare prodotti in tutte le aree scientifiche; di conseguenza ciascun ateneo ha selezionato i prodotti nelle aree in cui godeva di maggior prestigio presso la comunità accademica nazionale e internazionale. Proprio per questo la Vtr non può essere considerata un esercizio di valutazione complessiva dell'intero sistema universitario, bensì un tentativo di valutare gli atenei per la loro capacità di contribuire all'eccellenza della produzione scientifica nazionale (Reale, 2013a).

I tipi di prodotto sottoposti a valutazione sono stati: libri e capitoli di libro, articoli su rivista, brevetti, progetti, composizioni, disegni, performance, mostre, esposizioni, manufatti e opere d'arte. Il compito più oneroso per le strutture è stata la selezione dei prodotti da inviare a valutazione. La procedura più adottata ha previsto una prima selezione interna al dipartimento; i prodotti scelti sono stati poi ulteriormente selezionati dai Car, Comitati di area nominati dall'ateneo, e infine inviati ai Cat, Comitati di ateneo (costituiti dai presidenti dei Car); dopo essere stata certificata dal Nucleo di valutazione, la lista dei prodotti veniva inviata ai panel di area competenti.

Ciascun panel delle 20 aree scientifico-disciplinari è stato composto da un minimo di 5 a un massino di 9 esperti. I membri dei panel di area sono stati scelti in totale autonomia dal Civr, il quale attraverso una *call for experts* ha avuto cura di selezionare studiosi di comprovata esperienza nel settore scientifico di riferimento, con un elevato profilo accademico nazionale e internazionale (Civr, 2006).

Ai membri del panel di area competente è spettato il compito di nominare i due revisori di ciascun prodotto, che è stato poi valutato secondo cinque criteri: qualità, rilevanza, originalità e innovazione, internazionalizzazione[4].

[3] Nonostante ciò, tutte le strutture di ricerca pubbliche e private convenzionate con il Ministero hanno dato il proprio assenso.

[4] Nel bando scritto dal Civr sulla Vtr 2001-2003 si legge che i revisori esprimono un giudizio di merito sui prodotti scientifici che sono chiamati a valutare, secondo i criteri della: qualità, rilevanza, originalità/innovazione, internazionalizzazione. Il bando rimanda poi al documento

Per la selezione dei revisori si è proceduto alla creazione di un database di esperti valutatori, con la partecipazione di tutta la comunità accademica che ha individuato i migliori esperti di ogni area scientifica. Nel complesso i diversi panel di area hanno fatto ricorso a 6.661 revisori che hanno valutato 17.329 prodotti. La procedura di valutazione è stata la *peer blind review*: ciascun prodotto è stato letto da due revisori anonimi e giudicato come eccellente, buono, accettabile, limitato. Nel caso di giudizio eccessivamente discordante tra i due revisori, il Civr ha previsto la possibilità di richiedere la valutazione di un terzo esperto.

Inoltre, i prodotti venivano ponderati in funzione delle affiliazioni dei loro autori; si è infatti voluto dare un peso maggiore al prodotto eccellente scritto da più autori afferenti, tutti o quasi, alla stessa struttura che ha presentato il prodotto, rispetto al prodotto eccellente scritto da più autori afferenti, per la maggior parte, a strutture scientifiche diverse dall'ateneo che lo ha presentato. Solo nel primo caso si può parlare di un'eccellenza che si traferisce dal prodotto all'ateneo o centro di ricerca, poiché gli autori del prodotto valutato eccellente erano, al momento della pubblicazione, in carica presso la struttura che ha inviato il prodotto a valutazione; nel secondo caso no (Civr 2003a).

Nelle linee guida della Vtr 2001-2003 si legge che questo esercizio di valutazione si pone anche l'obiettivo di valutare la capacità delle strutture di sostenere e incentivare la ricerca nel medio periodo. Quattro sono stati gli indicatori utilizzati: numero di ricercatori in formazione (numero di dottorandi, assegnisti e borsisti di ricerca); il grado di mobilità internazionale in entrata e in uscita dei ricercatori di ruolo; l'ammontare delle risorse per la ricerca derivanti da fondi esterni all'ateneo (progetti Prin, fondi europei; ecc.); l'ammontare delle risorse interne destinate all'attività di ricerca.

Le strutture hanno poi ricevuto una valutazione sulla capacità di perseguire gli obiettivi di terza missione. Come indicatori sono stati

redatto dal Civr (2003b), contenente le linee guida per la valutazione della ricerca, per una trattazione più approfondita dei criteri di valutazione. Purtroppo, a un'attenta lettura del documento in questione, non si trova una chiara definizione dei quattro indicatori. Si legge infatti: «la valutazione della qualità scientifica e della rilevanza dei risultati si fonda sulla *peer review* e sull'applicazione di indicatori oggettivi, tra i quali, nei settori pertinenti, sono inclusi gli indici bibliometrici (in particolare, impact factor e citation analysis)». In nota si ricorda come «Il giudizio deve comprendere anche i riferimenti a originalità, innovazione e internazionalizzazione».

 ISBN 9788835103219

considerati il numero di brevetti depositati e/o attivi, e i costi e ricavi derivanti da ciascun brevetto[5].

Il 26 gennaio del 2006 sono stati diffusi tutti i giudizi sintetici dei due revisori di ciascun prodotto, e la *ranking list*, cioè la classifica delle strutture in base al giudizio medio dei propri prodotti inviati a valutazione[6]. I panel hanno redatto anche un documento con i punti di forza e le criticità di ciascuna area scientifica (Civr, 2006). Ogni ateneo ha potuto così controllare la valutazione dei prodotti inviati, la propria posizione in classifica e le aree scientifiche in cui ha ricevuto le valutazioni migliori (Bressan, 2007).

La Vtr 2001-2003 è considerata un punto di riferimento per i successivi esercizi di valutazione per via di alcune pratiche di qualità, tra cui il metodo di valutazione dei pari, che ha evitato il ricorso a sterili coefficienti bibliometrici (specialmente nelle aree umanistiche dove questi sono poco diffusi), e l'efficiente sistema informatico che ha supportato gli atenei in tutte le fasi di selezione dei prodotti (al riguardo vedi Frabboni e Sacchetta, 2005). D'altronde, non sono mancate le critiche, tra cui la libertà concessa agli atenei di scegliere in quali aree scientifiche inviare i prodotti meglio valutabili, tralasciando le aree in cui gli atenei erano meno attrezzati per una rigorosa valutazione esterna, e quella di selezionare più prodotti pubblicati da un solo ricercatore, particolarmente brillante, trascurando i docenti meno attivi. Questa strategia ha reso la Vtr un esercizio di valutazione della produzione scientifica di eccellenza e non di valutazione complessiva della ricerca del sistema universitario italiano.

In realtà l'impatto più consistente prodotto da questa valutazione è stato di tipo culturale, perché ha reso evidente che era non solo concretamente possibile ma anche opportuno un esercizio di valutazione della ricerca nelle università italiane, obiettivo considerato fino allora non fattibile e non desiderabile da larga parte del corpo accademico. In diversi settori disciplinari le comunità accademiche si sono trovate a dover riflettere su un esito che rendeva conoscibile a tutti la qualità attribuita alle proprie università rispetto ad altre sedi accademiche, innestando un dibattito che è stato molto formativo rispetto all'importanza che lo strumento valutativo

[5] Questo indicatore è stato aspramente criticato, vista la mancanza di dati ufficiali sui costi e ricavi di ciascun ateneo in merito ai propri brevetti, spin-off, ecc.

[6] Le strutture sono state differenziate in base alle proprie dimensioni: piccole, se hanno inviato da 1 a 9 prodotti; medie, quando hanno presentato da 10 a 24 prodotti; grandi da 25-74; mega, oltre i 75 prodotti.

può avere per l'organizzazione e il governo interno dell'università. Sotto questo profilo, la Vtr ha rappresentato un punto di non ritorno, consolidando in via definitiva l'idea che la valutazione è attività possibile e ineludibile, anche quando si tratti di valutazione della ricerca (Reale, 2013b, p. 152).

1.4. La valutazione della qualità della ricerca, la Vqr 2004-2010

La Vqr 2004-2010 è il secondo esercizio di valutazione nazionale della qualità della ricerca prodotta dal sistema universitario italiano. Con il d.m. n. 17 del 2011 il Miur incarica l'Anvur, la neonata Agenzia nazionale di valutazione del sistema universitario e della ricerca, della scrittura del bando della Vqr 2004-2010 e dell'organizzazione dell'intero esercizio di valutazione.

Con la legge n. 286 del 2006 l'Anvur sostituisce il Civr e il Cnvsu ereditandone il ruolo e le funzioni, con molteplici aggiunte che prevedono l'analisi e la valutazione complessiva dell'intero sistema universitario italiano[7]. Soprattutto, l'Anvur è l'organismo deputato all'organizzazione del secondo esercizio di valutazione nazionale della qualità della ricerca prodotta dal sistema universitario italiano, e, come detto nel par. 1.1, gli esiti di tale valutazione hanno conseguenze sulla distribuzione premiale dei fondi.

Rispetto all'esercizio di valutazione che l'ha preceduta, la Vqr 2004-2010 ha presentato numerose novità. In primo luogo, cambia l'arco temporale in cui sono stati pubblicati i prodotti sottoposti a valutazione: si passa da tre a sette anni, per rispondere alle critiche di gran parte della comunità accademica che in seguito alla Vtr aveva obiettato come tre anni fossero troppo pochi per raccogliere un paniere di prodotti sufficientemente rappresentativo della qualità espressa da un ateneo (Minelli Rebora e Turri, 2008). Inoltre, cambia l'unità della valutazione, non più il solo ateneo bensì l'ateneo e suoi dipartimenti considerati come l'unità di funzionamento essenziale per il modello organizzativo dell'università; un cambiamento che sposta l'occhio della valutazione dai livelli più alti alle strutture intermedie di *governance* dell'università.

I tipi di prodotto sottoposti a valutazione sono gli stessi della Vtr: libri e capitoli di libro, atti di congresso, articoli su rivista, brevetti,

[7] Vedi il par. 1.6 per un approfondimento sull'Anvur.

composizioni, disegni, performance, mostre, esposizioni, manufatti e opere d'arte, banche dati e *software*, carte tematiche.

A ciascun docente di ruolo è stato chiesto di inviare 3 prodotti pubblicati dal 2004 al 2010; ciò ha comportato un'impennata del numero di prodotti inviati a valutazione: si passa dai 17.329 della Vtr ai 179.280 della Vqr. Tutti i docenti hanno dovuto selezionare e inviare i propri prodotti, nessuno escluso, impedendo quindi all'ateneo di sceglierne una quota maggiore tra i docenti reputati più brillanti. Il docente che non ha inviato alcun prodotto è stato considerato "non attivo" e chi ne ha inviati meno di 3 "parzialmente attivo"; i docenti non attivi o parzialmente attivi hanno causato una penalizzazione nel giudizio finale del proprio dipartimento e ateneo.

Il concetto di "ricercatore attivo" introduce un parametro che valorizza quanti effettivamente contribuiscono all'attività di ricerca, e deprime chi non contribuisce, creando una differenziazione di status all'interno del corpo accademico. Il numero di pubblicazioni da sottomettere diventa indirettamente un parametro minimo di produttività per i singoli ricercatori: la Vqr propone, infatti, un numero di lavori di alta qualità che è lecito attendersi da ciascun ricercatore o docente di qualunque disciplina o settore (Reale, 2013b, p. 153).

Il bando Vqr chiarisce che i docenti devono fare ricorso allo strumento informatico messo a disposizione dal Cineca per indicare, in ordine di priorità, i prodotti di ricerca dai quali l'università di appartenenza sceglierà poi quelli da inviare a valutazione; alla struttura spetta l'importante compito di dirimere i conflitti di attribuzione in caso di coautoraggio, poiché un prodotto può essere attribuito a un solo autore[8]. Rispetto alla Vtr 2001-2003 i docenti assumono un ruolo più rilevante nel processo complessivo, poiché l'università può solo scegliere i prodotti da inviare ai Gev tra quelli presenti nella lista e selezionati *in primis* dai ricercatori (Rebora e Turri, 2010).

Per ogni area scientifica[9] l'Anvur nomina i Gev, i gruppi di esperti

[8] I prodotti inviati due volte dal medesimo ateneo sono valutati una volta sola, anche se scritti da due o più autori afferenti allo stesso ateneo; se invece un prodotto è stato scritto da più autori di diverse università, quello è valutato più volte.

[9] Rispetto alla Vtr, nella Vqr è cambiato il numero delle aree scientifiche: non più 20 bensì 14. Rimane invariata la possibilità che, su richiesta dei Gev, siano istituite sub-aree per le aree caratterizzate da particolare eterogeneità disciplinare.

 ISBN 9788835103219

valutatori, che a loro volta nominano i due revisori di ciascun prodotto. I Gev di area ricevono il prodotto in formato elettronico, corredato da alcuni metadati bibliografici, tra i quali il settore scientifico di riferimento. I prodotti afferenti ad alcune aree disciplinari sono stati valutati attraverso specifici indici bibliometrici[10]; si tratta della novità più rilevante rispetto alla Vtr, che invece si è basata esclusivamente sulla *peer review*. Ai Gev è stato affidato il compito di decidere quali prodotti vanno valutati con l'uno o l'altro approccio, in base al tipo di prodotto e all'area scientifica dello stesso[11], con l'unico limite di mantenere prevalente il numero di prodotti valutati con *peer review.*

Il giudizio dei revisori è stato articolato sui criteri della rilevanza, originalità, internazionalizzazione. Nel bando dell'Anvur si legge che per rilevanza «è da intendersi il valore aggiunto per l'avanzamento della conoscenza nel settore e per la scienza in generale, anche in termini di congruità, efficacia, tempestività e durata delle ricadute», per originalità «il contributo all'avanzamento di nuove conoscenze o nuove acquisizioni nel settore di riferimento», per internazionalizzazione «il posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale, e apprezzamento della comunità scientifica» (Anvur, 2011, p. 7).

Lo strumento di lavoro dei revisori è la scheda di valutazione, composta da tre domande chiuse, ognuna delle quali rileva uno dei criteri di valutazione della qualità del prodotto. La somma dei punteggi equivalenti ai giudizi espressi dai revisori su ciascuno dei tre indicatori costituisce il giudizio sintetico di qualità, che assegna il prodotto a una delle classi di merito[12]. Rispetto alla Vtr, cambia il numero delle classi e il punteggio di ciascun prodotto che in esse è classificato. Nella Vtr il giudizio era articolato su quattro livelli: eccellente (1), il prodotto si colloca tipicamente nel 20% superiore della scala di valore condivisa dalla comunità scientifica internazionale; buono (0,8), il prodotto si

[10] In particolare, si fa riferimento all'analisi del numero di citazioni del prodotto e al fattore di impatto della rivista, ove applicabile.

[11] L'indicazione data ai Gev dall'Anvur è di prendere in considerazione la valutazione bibliometrica solo per i tipi di prodotto (come articoli su rivista o proceedings) per i quali esistono già basi di dati adatti all'utilizzo di tali metriche. Inoltre i Gev hanno dovuto tenere in debito conto il settore scientifico disciplinare del prodotto; è risaputo che nelle discipline scientifiche il ricorso alla bibliometria è pratica da tempo diffusa; nelle scienze umanistiche non lo è ancora.

[12] Alla fine della compilazione della scheda, al revisore appariva la classe di merito finale di assegnazione che poteva eventualmente modificare, cambiando i punteggi dati ai tre indicatori.

colloca nel segmento 60% - 80%; accettabile (0,6), il prodotto si colloca nel segmento 40% - 60%; limitato (0,2) il prodotto si colloca nel 40% inferiore. Nella Vqr, invece, le classi di merito sono sei: eccellente (1), la pubblicazione si colloca nel 20% superiore della scala di valore condivisa dalla comunità scientifica internazionale; buono (0,8), la pubblicazione si colloca nel segmento 60% - 80%; accettabile (0,5), la pubblicazione si colloca nel segmento 50% - 60%; limitato (0), la pubblicazione si colloca nel 50% inferiore; la pubblicazione non è valutabile (-1); la pubblicazione non è stata presentata (-0,5).

Dall'analisi delle nuove classi di merito e dei relativi punteggi si evince quanto nella Vqr, rispetto alla Vtr, sia stata incrementata la selettività nel giudicare un prodotto di qualità. In primo luogo, i prodotti che ricadono nella fascia peggio valutata non ottengono alcun punteggio; nella Vtr avevano un peso di 0,2. Inoltre, i prodotti accettabili sono valutati 0,5 e non più 0,6 ma soprattutto aumenta la distanza, in termini di valutazione, tra questi e i prodotti considerati buoni ed eccellenti. Infine, sono istituite due classi di merito che danneggiano gravemente i ricercatori inattivi e gli atenei che, per errore o incuria, hanno inviato prodotti non valutabili (Rebora e Turri, 2010).

Nel complesso la Vqr 2004-2010 si è rivelata essere un esercizio di valutazione di enorme portata, che ha coinvolto 95 università e 38 enti di ricerca, 61.822 ricercatori, 436 esperti valutatori (d'ora in poi Ev) che hanno composto i Gev, 14.770 revisori di cui 4.620 in servizio presso strutture straniere. Inoltre, per la prima volta la connessione tra l'esito della valutazione e l'allocazione dei fondi pubblici all'università diventa uno dei cardini dell'intero esercizio di valutazione, a tal punto che le relazioni redatte dall'Anvur e inviate al Miur sono state poi utilizzate per ripartire quote considerevoli del finanziamento pubblico alla ricerca (Rebora, 2010).

1.5. La valutazione della qualità della ricerca, la Vqr 2011-2014

La Vqr 2011-2014 è il terzo esercizio di valutazione del sistema universitario italiano, istituito dal Miur che emana il d.m. n. 458 del 2015 contenente le linee guida. L'Anvur è l'agenzia incaricata di organizzare l'intero esercizio di valutazione e di scrivere il bando che sarà poi pubblicato il 30 luglio 2015.

 ISBN 9788835103219

La Vqr 2011-2014 è caratterizzata da alcuni elementi di discontinuità rispetto alla Vqr precedente. La novità principale sta nell'ampiezza dell'arco temporale di 4 anni che viene considerato utile per selezionare i prodotti pubblicati da sottoporre a valutazione; tale periodo era di 7 anni nella precedente Vqr e di 3 nella prima Vtr. Si ha quindi la sensazione che questo nuovo *range* costituisca un parametro stabile nei prossimi esercizi di valutazione.

I tipi di prodotto sottoposti a valutazione rimangono sostanzialmente gli stessi: libri e capitoli di libro, articoli su rivista, brevetti, composizioni, disegni, performance, mostre, esposizioni, manufatti e opere d'arte, banche dati e *software*, carte tematiche (Anvur, 2015). Questa volta vengono però esclusi i manuali e testi meramente didattici, le recensioni di un singolo lavoro che non abbiano analisi critica della letteratura sull'argomento, brevi voci enciclopediche o di dizionario senza carattere di originalità, brevi note a sentenza di tipo redazionale senza carattere di originalità o meramente ricognitive, brevi schede di catalogo prive di contributi scientifici autonomi.

Il numero di prodotti da inviare a valutazione passa da 3 a 2 per ciascun docente universitario; cambia anche il numero delle aree scientifiche, che passano da 14 a 16[13]. La nomina dei Gev di ciascuna area rimane prerogativa dell'Anvur; come nella precedente Vqr, compito principale dei Gev è nominare i revisori dei prodotti: due per ciascuno di essi[14].

Come nella precedente Vqr, anche per questo esercizio di valutazione lo strumento di lavoro dei revisori è la scheda di valutazione. In parte cambiano i criteri di valutazione della qualità dei prodotti: originalità, rigore metodologico, impatto nella comunità scientifica internazionale di riferimento. A seguito dei giudizi formulati dai revisori su queste tre dimensioni, ogni pubblicazione è attribuita a una delle seguenti classi di merito: eccellente, elevato, discreto, accettabile, limitato, non valutabile[15].

Dall'analisi delle nuove classi di merito e i relativi punteggi si evince quanto nella seconda Vqr sia stata incrementata la selettività nel giudicare un prodotto di qualità eccellente: sono considerati eccellenti

[13] Vedi la Tab. 1 e i successivi commenti sulle ragioni di tale ampliamento.

[14] Sui criteri di composizione dei Gev e di selezione dei revisori dei prodotti vedi il cap. 3.

[15] La scheda di valutazione dei revisori, i criteri di giudizio della qualità del prodotto e le classi di merito cui il prodotto è attribuito sono descritti analiticamente nel capitolo 4.

i prodotti che idealmente si trovano nel primo decile della produzione scientifica e non più nel primo 20%. In compenso i punteggi vengono distribuiti più equamente tra le classi cui sono attribuiti i prodotti di qualità intermedia. Nella prima Vqr i prodotti di qualità buona, "elevata" nella seconda Vqr, avevano un punteggio di 0,8; nella seconda Vqr di 0,7. Nella Vqr 2004-2010 i prodotti accettabili, "discreti" nella Vqr 2011-2014, avevano 0,5; nella seconda 0,4. Al resto dei prodotti di qualità inferiore è assegnato un punteggio di 0,1 e non più 0. Cambia anche il punteggio (0) assegnato ai prodotti attribuiti alla classe "non valutabile", che quindi non contribuiscono più a deprimere il punteggio generale dell'istituzione di appartenenza (nella prima Vqr ricevevano un punteggio di -1).

Non mancano gli elementi di continuità tra i due ultimi esercizi di valutazione. Le tecniche di valutazione rimangono invariate. I prodotti sono stati valutati sia con informazioni di natura bibliometrica (utilizzando soprattutto l'*impact factor* della rivista su cui è pubblicato il prodotto e il numero di citazioni ricevute dal prodotto[16]), sia con l'*informed peer review*[17], ricorrendo alle informazioni associate a ciascun prodotto inviato a valutazione (i metadati bibliografici del prodotto, identificativo Orcid dell'autore o dei coautori, l'area e il settore concorsuale e il settore scientifico disciplinare, l'*abstract* del prodotto, la lingua del prodotto, l'indicazione che il prodotto viene da un'area scientifica meglio valutabile con *peer review* o con indici bibliometrici).

Nella Tab. 1 sono riportati i dati relativi ai prodotti valutati con *peer review* o con bibliometria nelle diverse aree scientifiche. Da essa risulta che i prodotti di area 8a, 10, 11a, 12 e 14 sono stati valutati esclusivamente tramite la *peer review*.

Questa tabella mostra anche come la disaggregazione dell'area 8 e dell'area 11 in quattro distinte aree sia stata quanto mai opportuna. Nella Vqr 2004-2010 architettura e ingegneria civile costituivano l'area 8; scienze storiche, filosofiche, psicologiche e pedagogiche erano parte

[16] Se i due indici bibliometrici hanno valutato il prodotto in modo diametralmente opposto (ad esempio, un prodotto con elevato numero di citazioni pubblicato su una rivista con impatto molto basso o viceversa), lo stesso è stato classificato IR e sottoposto a valutazione con *peer review*.

[17] Per comporre l'albo dei revisori, incaricati di valutare i prodotti con *peer review*, l'Anvur ha selezionato alcuni nomi dal database Reprise del Miur, con criteri di merito scientifico (indice H di Hirsch, numero di citazioni, produzione scientifica recente), e altri non inclusi in Reprise, valutati con i medesimi criteri. Solo il Gev 12 ha pubblicato un modulo di autocandidatura per chi avesse voluto dare il proprio contributo come revisore. Il database finale è di 14.000 nomi.

dell'area 11. Vista l'eccessiva eterogeneità tra le discipline presenti all'interno di ciascuna delle due aree, nella Vqr 2011-2014 si è deciso di scorporare architettura (la nuova area 8a), valutata esclusivamente con *peer review*, da ingegneria civile (ora area 8b), valutata principalmente con indici bibliometrici, e le scienze storiche, filosofiche e pedagogiche (la nuova area 11a), valutate con giudizio dei pari, dalle scienze psicologiche (ora area 11b), valutate soprattutto bibliometricamente.

Tab. 1 - Numerosità e percentuale di prodotti valutati con peer review *o con bibliometria nelle diverse aree scientifiche*

Area	*Prodotti conferiti*	*Prodotti sottoposti alla* peer review	*% sui prodotti conferiti*	*Prodotti con valutazione bibliometrica*	*% sui prodotti conferiti*	*Prodotti con valutazione IR*	*% sui prodotti conferiti*
1	6.062	2.356	38,9	3.680	60,7	1.074	17,7
2	10.588	1.291	12,2	9.287	87,7	890	8,4
3	6.897	1.394	20,3	5.464	79,2	1.049	15,2
4	4.430	1.257	28,4	3.121	70,5	840	19,0
5	10.986	2.183	19,9	8.672	78,9	1.758	16
6	16.693	3.731	22,4	12.722	76,2	2.266	13,6
7	7.541	2.463	32,7	5.040	66,8	1.343	17,8
8a	3.456	3.433	99,3	5	0,1	0	0
8b	2.832	996	35,2	1.830	64,6	543	19,2
9	11.564	3.346	28,9	8.142	70,4	1.887	16,3
10	8.744	8.720	99,7	5	0,1	3	0
11a	6.123	5.956	97,3	151	2,5	37	0,6
11b	2.276	868	38,1	1.380	60,6	397	17,4
12	8.488	8.431	99,3	0	0	0	0
13	8.385	2.662	31,7	5.531	65,9	15	0
14	2.971	2.953	99,4	9	0,3	1	0,3
Totale	118.036	52.040	44,1	65.039	55,1	12.103	10,3

Fonte: Anvur (2017)

La valutazione dei prodotti di area 2, 3, 4, 5, 6, 7, 8b, 9 e 11b è stata in larga misura bibliometrica e ha riguardato gli articoli pubblicati su riviste indicizzate nei database di Wos[18] e Scopus[19]. L'Anvur ha acquistato da Thomson-Reuters ed Elsevier le informazioni bibliometriche di tali archivi per gli anni 2011-2014. Nel momento dell'invio dei prodotti a valutazione, alle università è stato chiesto di specificare la base dati (WoS oppure Scopus) adatta alla valutazione bibliometrica di ciascun prodotto,

[18] login.webofknowledge.com
[19] www.scopus.com

e l'indicatore di impatto preferito (IF5Y, Article Influence Score per WoS e IPP, SJR per Scopus). Oltre alle basi di dati WoS e Scopus, il Gev 1 ha utilizzato MathSciNet dell'American Mathematical Society, limitatamente all'indicatore d'impatto della rivista. Il Gev 13 ha impiegato un algoritmo bibliometrico che valorizza i prodotti con un numero di citazioni significativamente alto e quelli pubblicati da autori "prestigiosi" (tale indicatore è la città dove l'editore del prodotto ha sede).

Come nella prima Vqr, anche nella seconda Vqr sono stati oggetto di valutazione le università e i dipartimenti e non i singoli docenti. Per ciascuna università e dipartimento è stato stilato un profilo di qualità complessivo che è la combinazione del profilo di qualità dei prodotti di ricerca inviati a valutazione (che pesa per il 95%) e della capacità dell'istituzione di attrarre finanziamenti internazionali e statali (per il restante 5%). Per stabilire il profilo di qualità dei prodotti di ricerca di ciascuna università si è calcolato: il rapporto tra la somma delle valutazioni attribuite ai prodotti dell'ateneo nell'area scientifica di riferimento e la valutazione complessiva di area (indice Iras1); il rapporto tra il voto medio attribuito ai prodotti dell'università nell'area di riferimento e il voto medio ricevuto da tutti i prodotti dell'area (indice R); il rapporto tra la frazione di prodotti valutati eccellenti dell'istituzione nell'area di riferimento e la frazione di prodotti eccellenti dell'area (indice X).

Nella Vqr 2004-2010, al contrario, il profilo di qualità dei prodotti dell'ateneo era stabilito sostanzialmente da un solo indice, esito della somma dei punteggi conseguiti in base alle valutazioni dei prodotti inviati. In compenso questo indice pesava solo per il 50% della valutazione complessiva dell'istituzione; per il restante 50% contribuivano altri indici sintetici legati alla capacità dell'ateneo di attrarre finanziamenti pubblici e privati e sostenere adeguatamente la formazione e la mobilità dei propri ricercatori.

Come la precedente Vqr, anche questa si è rivelata essere un esercizio di valutazione di enorme portata, che ha coinvolto 96 università e 18 enti di ricerca, 60.455 addetti alla ricerca, 436 Ev che hanno composto i Gev (sulla distribuzione dei Gev tra le diverse aree scientifiche vedi la Tab. 2), 16.969 revisori di cui 3.423 in servizio presso strutture straniere[20].

[20] Nell'area 1 e nell'area 9 i revisori stranieri sono circa il 60% del totale, mentre nelle altre aree prevalgono i revisori italiani.

 ISBN 9788835103219

Tab. 2 - Numerosità dei Gev nelle diverse aree scientifiche

Area scientifica	*Numerosità dei Gev*
1 - Scienze matematiche e informatiche	22
2 - Scienze fisiche	33
3 - Scienze chimiche	23
4 - Scienze della terra	15
5 - Scienze biologiche	33
6 - Scienze mediche	62
7 - Scienze agrarie e veterinarie	23
8a – Architettura	16
8b - Ingegneria civile	11
9 - Ingegneria industriale e dell'informazione	38
10 - Scienze dell'antichità, filologico-letterarie e storico-artistiche	43
11a - Scienze storiche, filosofiche e pedagogiche	29
11b - Scienze psicologiche	6
12 - Scienze giuridiche	39
13 - Scienze economiche e statistiche	31
14 - Scienze politiche e sociali	12
Totale	436

Fonte: Anvur (2017)

I prodotti inviati a valutazione sono 118.036, di cui il 73,5% sono articoli su rivista, il 19,9% sono monografie/contributi su libro/curatele, il 5,8% contributi e *abstract* in atti di convegno (0,8% altro). Il 76,6% dei prodotti inviati è in lingua inglese; tale percentuale supera il 90% nelle aree bibliometriche; nelle aree non bibliometriche delle scienze umane, giuridiche e sociali prevale invece la lingua italiana.

Nel complesso il sistema universitario italiano, e in generale il mondo della ricerca, è stato valutato positivamente (vedi Tab. 3). Il 32,6% dei prodotti scientifici è valutato dai revisori eccellente (A) perché raggiunge i massimi livelli in termini di originalità e rigore metodologico, e ha conseguito o è presumibile che consegua un forte impatto nazionale e internazionale. Non basta. Il 30,8% dei prodotti è considerato di qualità elevata (B) avendo buoni livelli in termini di originalità e rigore metodologico, con un impatto significativo nella comunità scientifica nazionale e internazionale. Insomma, quasi i due

terzi della produzione scientifica italiana sono di alta qualità. Solamente l'11,6% è considerato accettabile (D) e il 3,5% limitato (E).

Tab. 3 - Numerosità e percentuale di prodotti attribuiti alle diverse classi di valutazione Vqr

	A	*B*	*C*	*D*	*E*	*F*	*Totale prodotti inviati*
Numero di prodotti inviati	38.435	36.394	24.433	13.639	4.178	957	118.036
% di prodotti inviati	32,6	30,8	20,7	11,6	3,5	0,8	100

Fonte: Anvur (2017)

Ovviamente non tutte le aree scientifiche contribuiscono egualmente a tale risultato (vedi Tab. 4).

Le scienze fisiche e le scienze chimiche sono le due aree scientifiche che hanno la quota maggiore di prodotti valutati come eccellenti. Seguono poi le scienze mediche, ingegneria industriale e dell'informazione, le scienze matematiche e informatiche, le scienze biologiche e ingegneria civile. Architettura, le scienze politiche e sociali e quelle giuridiche contribuiscono poco alla produzione scientifica di qualità eccellente.

Tab. 4 - Percentuale di prodotti per area attribuiti alle diverse classi di valutazione Vqr

	Area scientifica	*A*	*B*	*C*	*D*	*E*	*F*
1	(Scienze matematiche e informatiche)	38,4	28,0	18,2	10,8	4,2	0,4
2	(Scienze fisiche)	62,2	21,6	10,4	4,7	0,9	0,1
3	(Scienze chimiche)	49,2	32,0	12,9	4,6	0,8	0,6
4	(Scienze della terra)	27,9	29,7	21,6	14,2	5,5	1,2
5	(Scienze biologiche)	37,3	31,5	19,0	9,3	1,8	1,2
6	(Scienze mediche)	39,5	25,8	17,8	11,9	3,6	1,4
7	(Scienze agrarie e veterinarie)	28,4	31,5	19,4	14,9	5,4	0,5
8a	(Architettura)	8,6	34,2	35,9	16,0	4,8	0,5
8b	(Ingegneria civile)	37,6	29,3	17,7	12,6	2,5	0,2
9	(Ingegneria industriale e dell'informazione	38,6	27,6	18,2	12,3	2,7	0,7
10	(Scienze dell'antichità, filologico-letterarie e storico-artistiche)	18,1	46,2	25,4	8,7	1,4	0,2
11a	(Scienze storiche, filosofiche e pedagogiche)	16,1	42,4	29,2	10,2	1,8	0,3
11b	(Scienze psicologiche)	30,8	23,4	19,1	18,7	6,8	1,2
12	(Scienze giuridiche)	7,8	41,2	35,9	12,2	2,2	0,7
13	(Scienze economiche e statistiche)	24,6	22,9	17,9	19,5	12,7	2,3
14	(Scienze politiche e sociali)	8,3	32,5	34,5	20,0	4,4	0,3

Fonte: Anvur (2017)

1.6. L'Anvur

Come anticipato nel par. 1.4, l'Anvur è istituita con la legge n. 286 del 2006 nel cui testo si legge: «Al fine di razionalizzare il sistema di valutazione della qualità delle attività delle università e degli enti di ricerca pubblici e privati destinatari di finanziamenti pubblici, nonché dell'efficienza ed efficacia dei programmi statali di finanziamento e di incentivazione delle attività di ricerca e di innovazione, è costituita l'Agenzia nazionale di valutazione del sistema universitario e della ricerca (Anvur), con personalità giuridica di diritto pubblico, che svolge le seguenti attribuzioni: a) valutazione esterna della qualità delle attività delle università e degli enti di ricerca pubblici e privati destinatari di finanziamenti pubblici, sulla base di un programma annuale approvato dal Ministro dell'università e della ricerca; b) indirizzo, coordinamento e vigilanza delle attività di valutazione demandate ai nuclei di valutazione interna degli atenei e degli enti di ricerca; c) valutazione dell'efficienza e dell'efficacia dei programmi statali di finanziamento e di incentivazione delle attività di ricerca e di innovazione».

Nonostante la centralità del ruolo assegnato all'Anvur nel nuovo sistema di valutazione dell'università, i lavori dell'agenzia sono stati avviati solo quattro anni più tardi, per via dei molti ritardi sull'emanazione del regolamento attuativo. Mi riferisco al decreto n. 76 del 2010 del Presidente della Repubblica, «Regolamento concernente la struttura ed il funzionamento dell'Agenzia nazionale di valutazione del sistema universitario e della ricerca (Anvur)», che insieme alla legge n. 240 del 2010 di riforma dell'università ne incrementa ulteriormente le competenze. Oggi l'Anvur, oltre a organizzare gli esercizi nazionali di valutazione della qualità della ricerca, accredita e valuta i dottorati di ricerca, i corsi di studio e le sedi universitarie; ha compiti in materia di valutazione della didattica e di terza missione; accredita gli enti privati che richiedono il riconoscimento dello Stato; monitora la *performance* amministrativa delle università e degli enti pubblici di ricerca; valuta le riviste scientifiche per le aree non bibliometriche ai fini dell'Abilitazione Scientifica Nazionale (Asn), e redige il Rapporto sullo stato del sistema universitario e della ricerca (Fantoni, 2015).

Dal confronto tra le mansioni attribuite all'Anvur dalla legge n. 286 del 2006 e quelle che l'agenzia attualmente svolge, si osserva un in-

cremento considerevole degli impegni a essa affidati. Il primo regolamento disciplinava il ruolo dell'Anvur (condurre la valutazione esterna del sistema universitario italiano, indirizzare e coordinare l'attività dei Nuclei di valutazione interna, valutare i programmi di finanziamento pubblico agli atenei) in quanto organo tecnico che avrebbe dovuto valutare principalmente l'attività di ricerca delle università. Ad oggi, l'Anvur è incaricata di valutare non solo la qualità della ricerca, ma anche la qualità della didattica, della formazione, degli obiettivi di terza missione, e dei servizi amministrativi che le università pubbliche erogano ai cittadini utenti.

L'accorpamento del Civr e Cnvsu nell'Anvur non è stata solo un'operazione di trasferimento di funzioni amministrative da un ente pubblico a un altro. L'Anvur ha assunto le funzioni della sola e unica agenzia nazionale di valutazione del sistema universitario e della ricerca, con cui tutti gli operatori del sistema universitario italiano devono necessariamente confrontarsi. Una centralità che al contempo richiede, per essere legittimata, piena autonomia nei confronti del suo principale portatore di interesse, il Miur, che stabilisce l'entità dei finanziamenti destinati ai singoli atenei e dipartimenti in base agli esiti degli esercizi di valutazione organizzati dall'Anvur.

L'autonomia dell'agenzia rispetto all'operato del Ministero è sancita dalla legge. All'articolo 1 comma 3 del decreto n. 76 del 2010 del Presidente della Repubblica, si legge che: «L'Agenzia ha personalità giuridica di diritto pubblico (…). È dotata di autonomia organizzativa, amministrativa e contabile, anche in deroga alle disposizioni sulla contabilità generale dello Stato». Peraltro, questa autonomia è una delle ragioni per cui l'Anvur ha ricevuto l'accreditamento internazionale da parte dell'ENQA (*European Association for Quality Assurance in Higher Education*)[21].

Anche il meccanismo di nomina dei sette componenti del consiglio direttivo è trasparente e rafforza l'autonomia dell'agenzia dalla politica. All'articolo 8 comma 3 del suddetto decreto del Presidente della Repubblica si legge che: «I componenti del Consiglio direttivo sono nominati con decreto del Presidente della Repubblica, su proposta del

[21] Come si legge negli *Standards and guidelines for quality assurance in the European Higher Education Area*: «Agencies should be independent and act autonomously. They should have full responsibility for their operations and the outcomes of those operations without third party influences» (Enqa, 2015, p. 22).

Ministro, sentite le competenti Commissioni parlamentari (…). Ai fini della proposta, il Ministro sceglie i componenti in un elenco composto da non meno di dieci e non più di quindici persone definito da un comitato di selezione appositamente costituito con decreto del Ministro. Il comitato di selezione è composto da cinque membri di alta qualificazione, designati, uno ciascuno, dal Ministro, dal Segretario generale dell'OCSE e dai Presidenti dell'Accademia dei Lincei, dell'*European Research Council* e del Consiglio nazionale degli studenti».

Inoltre, il legislatore ha ritenuto opportuno istituire un Comitato consultivo, nominato dal presidente dell'Anvur su proposta del Consiglio direttivo, con la funzione di dare «pareri e formulare proposte al Consiglio direttivo, in particolare sui programmi di attività e sui documenti riguardanti la scelta dei criteri e dei metodi di valutazione», come si legge nell'articolo 11 del decreto n. 76 del 2010 del Presidente della Repubblica. I candidati a divenire i diciotto membri del Comitato consultivo sono indicati dal Consiglio Universitario Nazionale (Cun), dalla Conferenza dei Rettori delle Università Italiane (Crui), dal Consiglio Nazionale degli Studenti Universitari (Cnsu), dalla Conferenza dei Presidenti degli enti pubblici di ricerca, dall'Accademia dei Lincei, dal Consiglio nazionale dell'economia e del lavoro (Cnel), dalla Conferenza unificata Stato-Regioni, dall'*European Research Council*, dall'*European University Association,* dalla *National Union of Students in Europe*, dal Convegno permanente dei Direttori Amministrativi e dirigenti delle Università italiane (CoDau), e dal Segretario generale dell'OCSE. Il Comitato consultivo nasce dall'esigenza di istituzionalizzare un organo, interno all'agenzia ma composto da personalità del mondo dell'università e della ricerca, con cui l'Anvur si confronta per ricevere consigli e proposte in merito alla sua azione valutativa.

1.7. Una comparazione dei principali sistemi europei di valutazione della qualità della ricerca

In Europa il primo esercizio di valutazione nazionale della qualità della ricerca (il Rae, *Research Assesment Exercise*) si è tenuto nel 1986 in Inghilterra. In 5 occasioni[22] il Rae ha valutato la qualità della

[22] Nel 1989, 1992, 1996, 2001, 2008 (Palumbo, 2013).

produzione scientifica in tutti i paesi della Gran Bretagna. Nel 2014 il Rae è stato sostituito dal Ref (*Research Excellence Framework*), il cui impianto è stato confermato per l'immediato futuro: nel 2021 si svolgerà la seconda edizione del Ref.

In Gran Bretagna l'obiettivo dichiarato degli esercizi di valutazione della ricerca è sempre stato diversificare l'allocazione delle risorse pubbliche destinate alle università in base alla qualità della ricerca prodotta da ciascuna di esse. Tali attività, di gestione dell'esercizio di valutazione e di redistribuzione delle risorse, sono condotte da un'agenzia in ciascuna nazione britannica, coordinate dalla neonata agenzia Ukri (*United Kingdom Research and Innovation*). In Inghilterra se ne occupa il Rec (*Research England Council*), in Scozia il Sfc (*Scottish Founding Council*), in Galles l'Hefcw (*Higher Education Funding Council for Wales*), e il Dfe in Irlanda del Nord (*Department for the Economy in Northern Ireland*)[23]. L'esercizio di valutazione incide su una quota parziale dell'intero ammontare di risorse destinate alle università pubbliche; sono i cosiddetti *recurrent fundings*, che vengono assegnati in base agli esiti degli esercizi di valutazione.

Sono valutati i dipartimenti e le università. Ciascun dipartimento fornisce informazioni riguardanti il personale attivo nella ricerca, i prodotti di ricerca (4 prodotti per ciascun ricercatore, valutati con *peer review*), i corsi di dottorato, le fonti di finanziamento alla ricerca, e anche una descrizione accurata dell'ambiente di ricerca; non è valutata la didattica. La valutazione è condotta dai membri dei panel operanti nelle diverse aree scientifiche[24]; oggetto della valutazione sono la produzione scientifica, l'ambiente di ricerca, e una serie di indicatori di prestigio della struttura. Nella valutazione finale, il giudizio sulla produzione scientifica pesa per il 70%, l'ambiente di ricerca per il 20% e gli indicatori di prestigio per il restante 10%. Ciascun dipartimento ha la libertà di selezionare i ricercatori da sottoporre a valutazione; soli-

[23] Sono tutti enti pubblici i cui Boards sono di nomina politica.

[24] Nel Rae erano previsti 15 panel di esperti e relativi sotto-panel (da 3 a 8 per ogni panel), per un totale di 67 unità di valutazione scientifiche, ognuna per una differente area scientifico-disciplinare. Nel Raf le unità scientifiche di valutazione sono diventate 36. I panel principali coordinano i propri sotto-panel. I presidenti dei panel principali sono nominati dal ministro competente; i membri dei panel e dei sotto-panel sono esperti accademici di riconosciuta competenza nel proprio ambito scientifico, nominati dai direttori delle agenzie (Otley, 2010).

tamente «il dipartimento decide di presentare pochi ricercatori di qualità eccellente» (Perotti, 2008, p. 95). Il giudizio finale è articolato su 4 livelli: 1) qualità leader a livello mondiale, 2) qualità di eccellenza internazionale, 3) qualità riconosciuta a livello internazionale, 4) qualità riconosciuta a livello nazionale.

Nel 2006 il governo inglese ha avviato uno studio per riformare il Rae, specialmente a causa dei costi elevati e sempre crescenti. Obiettivo dichiarato del governo era passare da un sistema di valutazione basato sulla *peer review*, e per questo finanziariamente oneroso, a una valutazione bibliometrica dei prodotti scientifici. In un documento del Ministero delle Finanze si legge: «*The Government has firm presumption that after 2008 RAE the system for assessing research quality and allocating quality-related research funding to Universities from the Department for Education and Skills will be mainly metrics-based*» (HM Treasury, 2006, p. 3).

Nel 2008 si è tenuto l'ultimo Rae, e nel 2007 e 2010 si sono svolte due consultazioni per testare la possibilità di abbandonare *in toto* il sistema di valutazione basato sul giudizio dei pari con uno interamente bibliometrico; è stato condotto anche un esercizio pilota. Al termine di tale simulazione, si è deciso che la bibliometria avrebbe avuto solamente il ruolo di coadiuvare il giudizio dei pari, in quanto *informed peer review*.

Le novità principali del passaggio dal Rae al Ref sono: la diminuzione del numero di panel (36 anziché 67) e il peso assegnato, in sede di valutazione, all'impatto socioeconomico, quindi non accademico, dell'attività di ricerca. I valutatori Ref tengono in forte considerazione i cosiddetti obiettivi di terza missione, cioè la capacità dell'ateneo di fare ricerca che abbia ricadute a breve e medio termine sulle condizioni economiche e sociali dei cittadini e delle imprese del territorio in cui opera; lo studio di caso è lo strumento cui ci si affida per valutarne l'impatto. Sul giudizio finale alla struttura, l'impatto socioeconomico della ricerca pesa per il 20%, l'ambiente di ricerca per il 15% e la qualità della produzione scientifica per il 65% (Ref, 2010)

Nonostante queste novità, il programma di valutazione britannico ha mantenuto intatte le sue peculiarità. È infatti un programma intrinsecamente meritocratico perché *performance-based*, poiché premia la qualità e riconosce risorse aggiuntive a chi ha dimostrato di esserne

all'altezza. È valutata solo la ricerca, non i programmi di insegnamento; la valutazione è rigorosamente *ex post* da parte di un panel di esperti con l'uso combinato di *peer review* e indicatori bibliometrici; la valutazione è esterna all'ateneo e nazionale; l'esercizio di valutazione è finalizzato alla redistribuzione dei finanziamenti pubblici; la connessione tra i risultati dell'attività di valutazione e la ripartizione dei fondi pubblici è diretta (Hicks, 2011).

Ovviamente un sistema di valutazione totalmente *performance-based*, come quello britannico, non presenta solo vantaggi ma anche diverse controindicazioni. In primo luogo, una valutazione meritocratica, che si affida soprattutto al giudizio dei pari e solo parzialmente alla bibliometria, è molto costosa, in termini di risorse economiche e umane; richiede quindi finanziamenti sostanziosi e tempi lunghi, oltre a una capillare collaborazione di tutto il mondo accademico. Inoltre, se le università vengono valutate per la qualità della loro ricerca, il pericolo maggiore è che investano meno nell'attività didattica, diventando dei veri e propri centri di ricerca e non più l'istituzione cui è demandato il compito di trasmettere il sapere ai cittadini di domani. È probabile anche che le università concentrino le proprie risorse nelle aree scientifiche meglio spendibili in fase di valutazione (vedi le scienze ingegneristiche e mediche), deprimendo ulteriormente altre discipline (ad esempio quelle umanistiche e artistiche) già gravemente sottofinanziate (Geuna e Martin, 2003).

In Olanda l'attività di valutazione della qualità della ricerca universitaria è affidata a tre organismi: la *Royal Netherlands Academy of Arts and Sciences* (Knaw), la *Netherlands Organization for Scientific Research* (Nwo), l'*Association of Universities in Netherlands* (Vsnu). La Knaw raggruppa fino a 220 studiosi con elevato profilo scientifico, che svolgono attività di consulenza nei confronti del governo centrale in materia di politiche riguardanti il sistema nazionale della ricerca; il compito principale dell'organismo è finanziare 19 grandi istituti di ricerca in base agli esiti dell'esercizio di valutazione. La Nwo è un'agenzia governativa composta da un presidente e quattro membri nominati direttamente dalla corona, che rimangono in carica per 5 anni. Ci sono otto sezioni scientifiche, in ognuna delle quali sono valutati i progetti e le richieste di finanziamento che vengono sottoposti direttamente dai singoli ricercatori e dai gruppi di ricerca. Inoltre, 9 istituti di ricerca sono valutati e finanziati direttamente da questa agenzia. Infine, c'è la

Vsnu, un'associazione di rappresentanza del mondo universitario, composta dalle maggiori università del paese. Il suo ruolo è promuovere gli interessi dell'accademia olandese nei confronti del governo centrale, negoziando le politiche riguardanti il mondo dell'università. Nel 2002 al Knaw, al Nwo e al Vsnu il governo ha affidato l'onere di definire i criteri e le procedure con cui avviare il successivo esercizio di valutazione; la partnership tra accademia e politica ha dato alla luce il Sep (*Standard Evaluation Protocol*) 2009-2015 (Palumbo, 2013).

Il Sep valuta ogni sei anni gli istituti di ricerca attraverso una procedura di valutazione integrata: interna ed esterna all'università. La valutazione interna è una sorta di autoriflessione critica, condotta dallo stesso ateneo sottoposto a valutazione, sulla qualità del proprio lavoro di ricerca, che include la preparazione di una dettagliata documentazione sulla propria produzione scientifica, sulle strutture e sui programmi di ricerca cui l'università partecipa; è inviata ai valutatori quattro settimane prima della loro visita *in loco.* La valutazione esterna è implementata tramite una visita all'istituto di un gruppo di Ev, inviati dal Ministero dell'Università e della Ricerca, che intervistano il direttore del dipartimento, alcuni ricercatori e componenti dello staff (anche i dottorandi) per controllare le informazioni contenute nel documento di valutazione interna e per acquisirne di nuove utili alla valutazione.

Gli Ev sono nominati dai *Boards* delle tre agenzie, tra i nominativi suggeriti dagli istituti valutati; i valutatori sono soprattutto stranieri, così da evitare conflitti di interesse, e la numerosità dei singoli gruppi può variare molto in base alle esigenze delle agenzie. Il rapporto finale di valutazione è inviato ai *Boards* delle tre agenzie, che esprimono un giudizio sui risultati dell'attività di valutazione; tutti i documenti sono pubblicamente consultabili. Nel periodo che intercorre tra due valutazioni esterne, ciascun istituto è tenuto a riflettere sugli esiti della precedente valutazione per poter essere meglio valutato nella successiva occasione. Ciascun valutatore dà un giudizio finale su una scala di 5 livelli (eccellente, molto buono, buono, soddisfacente, insoddisfacente) che è poi discussa collegialmente per trovare un accordo comune (Knaw e Nwo e Vsnu, 2009).

Il sistema olandese non può quindi definirsi *performance-based* perché la valutazione non è condotta esclusivamente *ex-post* da valutatori esterni ma anche *ex ante* da Nuclei interni all'università. Inoltre, oggetto di valutazione non è la sola produzione scientifica, valutata

con indicatori bibliometrici o con *peer review*. Sul giudizio finale pesa molto la qualità dei programmi di ricerca (intesa come originalità delle idee e dell'approccio alla ricerca, come reputazione e profilo scientifico dei partecipanti al programma di ricerca), la qualità dei programmi di formazione dottorale (intesa come l'ammontare delle risorse destinate ai progetti di formazione dei giovani ricercatori e come integrazione dell'attività dottorale con i progetti di ricerca dell'istituto), l'impatto dell'attività di ricerca sulla comunità che vive nel territorio in cui opera l'ateneo (intesa come la capacità della ricerca di migliorarne le condizioni socioeconomiche), la qualità delle strutture di ricerca e la pianificazione degli investimenti futuri.

Le università, affrancate dagli oneri imposti da una valutazione *performance-based*, sono dunque libere di orientare le proprie politiche accademiche in base alla *mission* che si sono date, ad esempio investendo in progetti di ricerca e formazione a lungo termine, anche in settori disciplinari nuovi o in aree scientifiche che al momento risultano essere poco attrattive per il mondo industriale. È un sistema di valutazione che promuove la pluralità degli approcci alla ricerca, incoraggia l'integrazione tra ricerca e alta formazione, facilita l'indipendenza dell'ateneo rispetto alle esigenze ondivaghe e passeggere dei mercati. Inoltre, il Sep non ha l'obiettivo di ripartire le risorse pubbliche in base alla *performance* scientifica dell'università; lo scopo è promuovere una cultura di auto-valutazione interna alle strutture, che, attraverso un meccanismo di confronto e apprendimento critico, incoraggi i ricercatori a contribuire attivamente alla costruzione di progetti di ricerca di qualità che accrescano il prestigio del dipartimento e dell'università per cui lavorano (Rip e Van der Meulen, 1995).

In Francia il sistema di valutazione della qualità della ricerca è molto simile a quello in vigore nei Paesi Bassi. Qui si è deciso di accorpare le funzioni di valutazione del sistema universitario, prima svolte da diversi organismi, in un'unica agenzia di valutazione, l'Aeres (*Agence d'evaluation de la recherche et de l'enseignement superior)*, oggi Hceres (*Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur*). L'agenzia è stata istituita nel 2007, ed è un'autorità amministrativa indipendente – ciò è sancito nel suo statuto – con lo scopo di valutare il sistema universitario con trasparenza, equità e affidabilità. È composta da un consiglio direttivo di 25 membri con un alto profilo nel mondo scientifico, nominati dall'agenzia per quattro anni, i quali hanno

potere di indirizzo sulle politiche di valutazione. Per ciascuna delle tre sezioni scientifiche (*Sciences humanes et sociales, Sciences dures, Sciences de la vie*) sono nominati un centinaio di delegati scientifici, responsabili delle procedure di valutazione nella propria sezione. Il ruolo chiave è occupato dalle migliaia di esperti, che, ciascuno nel dominio scientifico di competenza, monitorano e valutano l'adeguatezza del metodo di valutazione rispetto al proprio ambito disciplinare; gli esperti sono formati al sistema di valutazione dell'agenzia e il loro *feedback* è cruciale per il miglioramento costante di tutte le procedure di valutazione (Mesr, 2007).

La differenza principale tra il sistema di valutazione francese e gli altri due finora analizzati è l'unità di valutazione: non solo le università e i dipartimenti, anche i singoli team di ricerca. Lo scopo del programma è individuare i migliori gruppi di ricerca – ogni anno ne sono valutati circa 700 – all'interno di ciascuna area scientifica, assegnare loro un punteggio e inserirli in una classifica nazionale.

La valutazione del sistema universitario francese si fonda su due pilastri: l'autovalutazione, attraverso la redazione di un rapporto di valutazione interna all'università, e la valutazione esterna, condotta dai valutatori dell'agenzia, che integra la prima. Il periodo di valutazione dura solitamente nove mesi[25] ed è articolato in tre fasi: preparazione, visita, restituzione. Nella prima fase il Nucleo di esperti valuta il rapporto di valutazione interno redatto dal singolo team di ricerca; nella fase successiva, i valutatori fanno visita agli evaluandi per raccogliere le informazioni di controllo rispetto a quanto dichiarato nel primo documento; infine, viene redatto il documento dell'attività di valutazione con le osservazioni dei valutatori. In questo lasso di tempo si avvia un dialogo costante tra l'agenzia e la struttura valutata, al punto che il rapporto finale non viene reso pubblico se prima se non si è data possibilità all'università di commentarlo e discuterlo ove si trovasse in disaccordo con le riflessioni dei valutatori (Mesr, 2009).

Il Hceres valuta soprattutto la ricerca, sia a livello universitario, dipartimentale e dei singoli gruppi di lavoro. L'esito del rapporto di valutazione tiene conto: della qualità della produzione scientifica tramite *peer review* e indici bibliometrici, della partecipazione a progetti e programmi

[25] Ciascuna struttura è valutata ogni quattro anni; ogni anno sono valutate le università che ricadono in una determinata area geografica.

di ricerca nazionali e internazionali, dei finanziamenti ottenuti da agenzie nazionali e internazionali, dell'apertura a nuovi ambiti di ricerca interdisciplinari (la cosiddetta ricerca di frontiera), degli investimenti nella diffusione della cultura scientifica (European Commission, 2010).

Affinché la produzione scientifica del quadriennio di riferimento sia considerata apprezzabile, un ricercatore deve avere pubblicato almeno due articoli o monografie di rango A, cioè che sono state referate da un comitato scientifico. Per articolare ancor di più la valutazione della produzione scientifica, le riviste scientifiche sono state classificate in tre classi (A, B e C), suddivise in base all'ampiezza della produzione scientifica in esse pubblicata. I gruppi meglio valutati e le università loro affiliate ricevono finanziamenti aggiuntivi dal Ministero dell'Istruzione e della Ricerca (Palumbo, 2013).

In Spagna la valutazione della ricerca universitaria nazionale è condotta dall'Anep (*Agencia Nacional de Evaluacion y Prospectiva*) e dal Cneai (*Comision Nacional Evaluadora de la Actividad Investigadora*). Questi due organismi non svolgono un vero e proprio esercizio di valutazione nazionale congiunto; ciascuno lavora con propri tempi e procedure.

L'Anep è l'agenzia del Ministero della scienza e dell'innovazione, nata con lo scopo di finanziare i progetti dei ricercatori che ne fanno richiesta. Per ciascuna area un team di 4-8 coordinatori nominati dal Sottosegretario di Stato alla ricerca su proposta del direttore dell'agenzia, supportati da 120 assistenti, seleziona da un database di 20.000 nomi i valutatori che giudicano in modo anonimo i progetti di ricerca. La valutazione è condotta *ex ante* tramite *peer review* da Ev nelle diverse aree disciplinari. Il progetto, per la cui realizzazione il ricercatore chiede il finanziamento, è inviato all'Anep e assegnato a una delle 26 aree tematiche; per ogni area tematica vi è un gruppo di coordinamento che seleziona i valutatori competenti nell'area di afferenza del progetto. Il progetto è valutato secondo le sue qualità intrinseche (significatività, realizzabilità, originalità e interdisciplinarità, chiarezza degli obiettivi, valenza dei risultati attesi) e la qualità della produzione scientifica dei suoi proponenti. Sulla scorta del giudizio di pari, si predispone la relazione finale e si comunica al ricercatore l'esito della valutazione (Palumbo, 2013).

Il Cneai, organo del Ministero, valuta la produzione scientifica di

quegli studiosi che decidono su base volontaria di sottoporsi a valutazione per ottenere uno stipendio più alto; l'incremento è a carico delle Regioni, e le commissioni sono nominate dal Ministero. La valutazione è dunque *ex post*. Il ricercatore chiede di sottoporre a valutazione la propria produzione scientifica; non c'è una selezione dei prodotti da parte del ricercatore, poiché la valutazione verte sull'intero *corpus* pubblicato negli ultimi sei anni. I prodotti sono assegnati a una delle 15 aree disciplinari previste; per ciascuna area è nominato un gruppo di coordinamento che rimane in carica per due anni. I valutatori, selezionati dal gruppo di coordinamento, dispongono di una serie di informazioni di carattere bibliometrico (grado di prestigio dell'editore o della rivista su cui si è pubblicato, il numero di citazioni ricevute, ecc.). Nelle scienze umane, dove indicatori di questo tipo sono poco utili, i valutatori si affidano alla *peer review* e ad altre informazioni, ad esempio la rigorosità delle procedure di revisione cui è stato sottoposto il prodotto poi pubblicato (Jimenez-Contreras, 2010).

La Germania è certamente il paese europeo con il sistema di valutazione più complesso, perché in esso coesistono modelli di valutazione differenti. Gli attori coinvolti nel processo di valutazione sono diversi; ciascuno con proprie procedure di valutazione, precipue tempistiche, finalità e referenti.

Il Dfg è l'ente nazionale di diritto privato, costituito da tutte le università e centri di ricerca tedeschi, che si occupa di distribuire il 70% dei fondi pubblici destinati alle università, stanziati dallo stato centrale e dai *Länder*. Il Dfg finanzia i progetti individuali dei ricercatori che ne fanno richiesta, le borse di studio ai dottorandi e giovani ricercatori per i loro studi di specializzazione, i finanziamenti a gruppi di ricerca (da 5 a 25 componenti) impegnati in progetti complessi che si protraggono anche per diversi anni. Il Dfg nomina i panel dei revisori dei progetti di cui sono chiamati a giudicare *ex ante* la qualità tecnico-scientifica (l'idea progettuale, l'entità del finanziamento richiesto, la rilevanza dei risultati attesi), il *know-how* di cui dispongono i proponenti e la coerenza del loro *curriculum vitae* e delle loro esperienze rispetto al progetto presentato. La valutazione è praticamente a ciclo continuo, e solitamente la procedura si conclude in breve tempo.

Oltre alla valutazione del Dfg, ciascun *Land*[26] gestisce autonomamente l'attività di valutazione esterna dei singoli atenei. Ad esempio, lo stato del Baden-Württemberg ha costituito un'agenzia *ad hoc*, la Evalag, cui ha affidato la valutazione delle attività scientifiche delle sue università. Il panel di Ev, solitamente ogni cinque anni, valuta l'intero ateneo, grazie a una serie di informazioni di cui dispone (qualità e numerosità della produzione scientifica, ammontare dei fondi raccolti su progetti internazionali, parametri relativi all'attività didattica). La procedura di valutazione è *ex post* con *informed peer review*. La produzione scientifica è valutata in base al numero e prestigio delle pubblicazioni degli addetti alla ricerca, all'impatto del prodotto (il numero di citazioni ricevute), al posizionamento nel *journal ranking*[27] della rivista in cui il prodotto è stato pubblicato. L'ultimo livello di valutazione è interno a ciascuna università, la cui finalità è redistribuire i fondi tra le singole facoltà, dipartimenti e ricercatori. Ciascuna università decide autonomamente quali procedure adottare e come costituire i Nuclei di valutazione interna. In alcuni atenei queste valutazioni sono informali; le quote di finanziamento sono negoziate in base a valutazioni scientifiche e didattiche (come il numero di studenti, le ore di lezione, numero di docenti). In altre università la negoziazione lascia spazio a procedure standardizzate e altamente formalizzate, specialmente nell'assegnazione dei fondi ai singoli ricercatori, considerando il profilo di *performance* scientifica del singolo studioso. L'attività di valutazione interna è solitamente annuale e posteriore all'assegnazione statale dei fondi (Palumbo, 2013).

La Svezia dal 2009 ha invece adottato un modello integralmente *performance-based*, in cui si valuta solo la produzione scientifica.

[26] L'attuale sistema di valutazione è centrato sull'autonomia dei Lander, principio ulteriormente rafforzato dall'ultima riforma costituzionale del 2006. Attualmente il governo federale contribuisce per il 55% allo stanziamento dei fondi per le università pubbliche; i Lander contribuiscono per il restante 45%.

[27] Il journal ranking è la classificazione delle riviste scientifiche nazionali e internazionali in base al loro prestigio presso la comunità accademica tedesca. Vista l'insufficienza di indici bibliometrici in grado di cogliere la qualità della maggior parte dei prodotti scientifici dei ricercatori tedeschi (Schmitz, 2008), l'Accademia Tedesca di Economia Aziendale ha avviato un lavoro di classificazione delle riviste scientifiche in base alla qualità del processo di revisione cui sono sottoposti gli articoli inviati a valutazione. Grazie a due survey nel 2002 e nel 2008, i ricercatori e professori hanno contribuito a differenziare le riviste scientifiche in 5 classi molto selettive; basti pensare che le riviste top ranking presenti nella prima classe (A+) sono appena 12.

L'attività di valutazione poggia su due pilastri: uso diffuso della bibliometria e ampliamento al 25% della quota premiale di fondi pubblici. Tale modello pesa poco sul bilancio dello Stato centrale perché la valutazione è esclusivamente bibliometrica e non prevede in nessun caso il ricorso al giudizio dei pari. Lo scopo è dare priorità alle realtà universitarie che sono state in grado di attirare in Svezia i ricercatori che fanno ricerca industriale di qualità. Di conseguenza, alcune discipline scientifiche hanno acquisito, più di altre, prestigio nel mondo accademico e riconoscibilità presso l'opinione pubblica. Molto finanziate e apprezzate sono le scienze ingegneristiche, mediche, le nano e bio-tecnologie, in grado di rispondere al meglio alle esigenze industriali del mercato e di contribuire al benessere generale del paese; le scienze umanistiche e sociali – ma anche altre scienze dure la cui ricerca non fornisce applicazioni immediatamente spendibili sul mercato – sono sempre più marginalizzate nella redistribuzione dei fondi pubblici e anche nella rappresentazione che l'opinione pubblica ha del rapporto tra scienza e società (European Commission, 2010).

 ISBN 9788835103219

2. La definizione del concetto di qualità della ricerca

di Antonio Fasanella

2.1. Introduzione

A prescindere dagli ambiti disciplinari di riferimento, la definizione di qualsiasi concetto scientifico è intrinsecamente *stipulativa*. Se, in questo modo, viene esautorata ogni velleità di ipostatizzazione, lo stabilimento di regole di ingaggio comuni fa sì che un uso determinato del concetto diventi ugualmente vincolante, tanto più in relazione alle finalità dell'uso stesso.

A seguito dei processi valutativi implicati dall'istituzione della Vqr, e in considerazione proprio degli obiettivi e delle conseguenze della stessa, una definizione vincolata e massimamente condivisa del concetto di qualità della ricerca risulta indispensabile. Questa prima parte del volume intende proporre un'analisi del concetto di qualità della ricerca per come esso è stato definito lessicalmente nell'ultima tornata valutativa (Vqr 2011-2014), mettendone in risalto gli aspetti problematici e suggerendo possibili soluzioni. Nel seguito, verranno considerate altresì le modalità applicative dello stesso concetto, analizzandone criticamente la definizione operativa senza peraltro abbandonare l'intento positivo, riformatore che caratterizza il nostro approccio.

Infatti, l'esercizio critico qui condotto si pone entro una prospettiva orientata in senso *progressivo*. In nessun modo si intende negare l'utilità o, se si vuole, la necessità di sottoporre a valutazione gli esiti della ricerca accademica. Si ribadisce tuttavia la necessità di un controllo pubblico dei processi di valutazione, perseguendo inflessibilmente un principio di *accountability* degli stessi, proprio al fine di far emergere

 ISBN 9788835103219

eventuali anomalie. Evidentemente, tali difetti possono minare la legittimità stessa della valutazione e, perciò, piuttosto che essere omessi o negati, se realmente esistenti richiedono di essere messi pienamente in luce, ridotti e, se possibile, rimossi, prima che possa farsi strada un orientamento di scetticismo e sfiducia verso la Vqr, verso questa Vqr. Inutile dire che tale orientamento, appunto, non necessariamente si traduce in un rifiuto della valutazione in quanto tale, ma certamente potrebbe favorire il ritorno a forme di valutazione alternative alla Vqr per come la conosciamo. Forme di valutazione che valorizzino gli aspetti di complessità dell'oggetto della valutazione stessa, la qualità della ricerca, che per sua natura sfuggirebbe a tentativi di standardizzazione e richiederebbe un approccio individuale, più qualitativo, capace di distinguere da caso a caso, attento alle specificità, ecc. Queste premesse aprono la strada a una valutazione evidentemente *judgemental*, basata su elementi di conoscenza tacita non interamente ricostruibili e su risorse squisitamente soggettive come l'intuizione, una certa sensibilità, una speciale attitudine, una spiccata capacità interpretativa del testo o del prodotto, la qualità del quale deve essere valutata; una valutazione, perciò, inevitabilmente caratterizzata da margini non troppo ampi di *accountability*.

Ebbene, tutta la trattazione offerta in questo volume si contrappone a una simile concezione della valutazione, prende atto positivamente degli sforzi che sono stati condotti in sede Vqr proprio al fine di superarla, ma non può non riconoscere, d'altra parte, che l'attuale Vqr, nonostante tali sforzi, è ancora contraddistinta da elementi di opacità, di ambiguità, di arbitrio, di scarsa ispezionabilità delle procedure, che la rendono suscettibile di essere perfezionata, anche sulla base delle proposte che saranno qui avanzate.

2.2. Determinatezza e uniformità d'uso del concetto di qualità

Possiamo assumere che il valore operativo di un concetto è dato dalla capacità di intercettare il più ampio consenso possibile. Nondimeno, si tratta di un principio accettabile mantenendo il discorso ad un livello di astrazione molto alto. In realtà, esiste sempre un delicato equilibrio fra l'ideale di massima estensione del concetto e la sua effettiva applicabilità, condizionata da fattori di carattere contestuale.

Così, le inevitabili specificità legate ai molteplici ambiti della conoscenza scientifica e, soprattutto, le numerose prospettive, talvolta apertamente in opposizione, che fanno capo a diverse scuole di pensiero in seno a un medesimo ambito disciplinare (Kuhn, 1962; tr. it., 1978), possono rendere molto difficile l'esercizio di definizione del concetto di qualità della ricerca. La difficoltà può riassumersi nella domanda: fino a che punto può essere estesa la nozione di qualità della ricerca? Ovvero, qual è il livello di specificità contestuale al quale deve potersi adattare?

In pratica, occorre far fronte a un problema liminare che l'Anvur ha inteso risolvere predisponendo una definizione iniziale, – per così dire – plastica del concetto di qualità della ricerca e delegando al Gruppo di Esperti Valutatori (Gev)[1] di ogni area scientifico-disciplinare il compito di redigere delle linee guida, al fine (1) di adattare la definizione alle specificità di area, (2) di renderla, così adeguata, pienamente operativa, senza che peraltro le due operazioni avessero potuto snaturare la sostanza del significato originario, generale di qualità della ricerca.

Se, da un lato, questo tipo di soluzione appare del tutto condivisibile, dall'altro sembra essere stata recepita solo parzialmente dai Gev. Limitando lo sguardo all'Area 14, ossia il gruppo disciplinare che d'ora in poi costituirà il dominio di riferimento del presente lavoro, si può osservare che le linee guida messe a punto dal competente Gev non sono in grado di adempiere ai due obiettivi sopra indicati. Infatti, come si può apprezzare dalla comparazione dei testi originali (Tab. 1), le linee guida rimandano circolarmente ai criteri di qualità della ricerca stabiliti dal Bando di istituzione della Vqr, senza offrire alcuna reale e accurata specificazione del significato del concetto di qualità nelle sue dimensioni costitutive (originalità, rigore e impatto). Non solo il testo delle linee guida aggiunge poco o nulla al contenuto del Bando, ma, paradossalmente, nella terza voce concernente l'impatto attestato o potenziale del prodotto è stato addirittura omesso un elemento del significato originario, ovvero il tipo di influenza, che può essere teorica e/o applicativa (ancora Tab. 1).

[1] Per un'analisi approfondita sul ruolo e la funzione dei Gev si rimanda al capitolo 3.

Tab. 1 - Formulazione del concetto di qualità della ricerca (bando e linee guida del Gev14)

	Bando	*Linee guida*
Originalità	da intendersi come il livello al quale il prodotto introduce un nuovo modo di pensare in relazione all'oggetto scientifico della ricerca, e si distingue così dagli approcci precedenti allo stesso oggetto	ha lo scopo di misurare quanto sia innovativo il prodotto di ricerca, rispetto a un nuovo modo di pensare, nuove prospettive, tesi, temi e/o fonti, in relazione all'oggetto scientifico della ricerca, e si distingue pertanto dai precedenti lavori sullo stesso tema
Rigore metodologico	da intendersi come il livello al quale il prodotto presenta in modo chiaro gli obiettivi della ricerca e lo stato dell'arte nella letteratura, adotta una metodologia appropriata all'oggetto della ricerca e dimostra che gli obiettivi sono stati raggiunti	ha lo scopo di misurare: i) il livello di chiarezza con cui il prodotto presenta gli obiettivi di ricerca; ii) il livello di competenza scientifica e di padronanza dello stato dell'arte; iii) la capacità di adottare una metodologia appropriata rispetto all'oggetto della ricerca; iv) il raggiungimento (realizzazione) degli obiettivi prefissi
Impatto attestato o potenziale nella comunità scientifica internazionale di riferimento	da intendersi come il livello al quale il prodotto ha esercitato, o è suscettibile di esercitare in futuro, un'influenza teorica e/o applicativa su tale comunità anche in base alla sua capacità di rispettare standard internazionali di qualità della ricerca	ha lo scopo di misurare il livello rispetto a cui il prodotto ha esercitato – o è presumibile eserciti in futuro – un'influenza su tale comunità, anche in base alla sua capacità di rispettare standard internazionali di qualità della ricerca.

Fonte: Anvur, 2015, p. 14; Anvur, 2017, p. 4

L'incursione senza le necessarie mediazioni della nozione di qualità della ricerca entro una comunità che né aveva contribuito alla sua generazione né si era interrogata a sufficienza circa le modalità della sua applicazione al settore specifico ha finito per rendere ancor più arduo il già delicato lavoro di valutazione dei prodotti assegnato ai *referees*[2]. In altri termini, è mancata un'azione, essenziale ai fini della valutazione, di precisazione del significato della formulazione originaria, in qualche modo necessariamente vaga al fine di essere applicata a settori scientifico-disciplinari distinti e non immediatamente omologabili. In tal senso, è necessario rimarcare il diffuso riferimento a molteplici dimensioni enucleate solo parzialmente, nonché l'utilizzo di locuzioni indeterminate e variamente interpretabili come, per esempio, il richiamo a non meglio precisati «standard internazionali di qualità

[2] Anche in questo caso si veda il capitolo 3 per una descrizione esaustiva di tale figura.

della ricerca» (Anvur, 2017, p. 4), o a un presunto «nuovo modo di pensare» (ibid.), che in quanto tale deve ancora costituirsi nelle sue specificità caratterizzanti e pertanto, in una prospettiva attuale, risulterebbe non ben definito, non chiaro.

In aggiunta vi è la controversa ubiquità temporale del giudizio relativo all'impatto attestato o potenziale, che in ipotesi dovrebbe essere formulato tenendo conto sia dell'effettiva influenza già esercitata sulla comunità scientifica di riferimento sia del presunto impatto futuro. È appena il caso di notare come una simile impostazione obblighi i revisori a mettere in gioco abilità predittive tutt'altro che scontate, di fatto incoraggiando il ricorso a parametri valoriali e disposizionali estremamente soggettivi. Richiedere una previsione, per giunta basata su non meglio definiti standard internazionali, pone ciascun revisore nelle condizioni di riferirsi a propri criteri di valutazione, con il rischio di privilegiare quei lavori che risultino più vicini ai propri orientamenti epistemologici, teorici e metodologici[3]. In estrema sintesi, definizioni concettuali lasche e opache rendono oltremodo oneroso il compito dei *referees*, costretti a un ingente lavoro ermeneutico. Tutto ciò, inoltre, aumenta la probabilità di introdurre nell'esercizio valutativo elementi distorsivi indesiderati.

In quest'ottica le linee guida dovrebbero svolgere un ruolo non sostituibile, ponendosi come il più importante baluardo a difesa del principio di *determinatezza* e *uniformità d'uso* (Hempel, 1952; tr. it., 1961, p. 13) del concetto di qualità della ricerca scientifica[4]. Sapendo che, venendo meno tale principio, vengono meno le condizioni di validità intersoggettiva e intrasoggettiva della valutazione; in altre parole, viene meno la garanzia stessa di comparabilità delle valutazioni espresse da valutatori diversi rispetto allo stesso prodotto della ricerca ovvero dallo stesso valutatore in merito a prodotti diversi.

Dunque, sarebbe stato opportuno mettere a punto delle vere e proprie *istruzioni per l'uso pratico del concetto di qualità della ricerca ai fini della valutazione*, dotate di un adeguato livello di articolazione e

[3] Si tenga presente che tali criticità non sono state disambiguate nemmeno nella scheda di valutazione finale - strumento operativo a disposizione dei *referees* per la valutazione dei contributi - come si potrà leggere nel capitolo 4, a cui si rinvia per un'analisi puntuale ed esaustiva dei bias legati ai problemi di formulazione dei criteri.

[4] È bene ricordare che l'obiettivo di uniformazione delle modalità d'uso implica un processo di confronto tra i diversi portatori di interesse: autori, revisori, esperti valutatori, decisori.

profondità, e perciò capaci di corrispondere alla veramente notevole varietà dei prodotti della ricerca, in assenza delle quali lo sforzo definitorio iniziale avrebbe rischiato di non risultare effettivo. In questo modo si sarebbero poste le condizioni reali per ottemperare al principio di determinatezza e uniformità d'uso del concetto da parte dei *referees*, riducendo al massimo il margine di interpretazione soggettiva sulla base di coordinate concettuali potenzialmente estranee rispetto al concetto in uso.

Si deve inoltre tener conto del fatto che un sistema di regole così concepito svolgerebbe una fondamentale funzione pedagogica dal punto di vista degli addetti (gli autori delle pubblicazioni), i quali potrebbero in questo modo tenere concretamente conto dei contenuti espressi in riferimento al concetto di qualità in modo che esso risulti chiaramente riproducibile nel lavoro di ricerca e nelle stesse pubblicazioni. Ciò ammettendo che il concetto di qualità della ricerca non sia suscettibile di decisivi cambiamenti tra una edizione e l'altra della Vqr, come pure è accaduto nel passaggio dalla prima alla seconda tornata valutativa (cfr. Di Benedetto, 2015. pp. 97-98).

Infine, sempre in una prospettiva formativa, va segnalato un problema di sincronizzazione. Diversamente da quanto accaduto, le linee guida andrebbero pubblicate ben prima dell'inizio della Vqr, costituendo, man mano, un riferimento stabile (o che tende alla stabilizzazione) a standard generali di qualità con cui gli addetti possono confrontare il proprio lavoro.

Senza indugiare oltre su ciò che avrebbe potuto essere e non è stato, si ritiene ora opportuno volgere lo sguardo in avanti, oltre i confini nazionali, cercando di comprendere se è possibile individuare delle buone pratiche analizzando l'operato di agenzie di valutazione della ricerca estere. Tra le esperienze delle nazioni europee che hanno messo in campo azioni valutative della ricerca scientifica, cui si è accennato nel capitolo 1, sicuramente interessanti risultano quella inglese e quella francese.

Il riferimento al Ref (*Research Excellent Framework*) è dovuto alla sua natura, potremmo dire, di "capofila" europeo nell'ambito della valutazione dei prodotti scientifici. Il suo primato non è esclusivamente

temporale[5], ma anche teorico e procedurale, viste la particolare accuratezza concettuale e le risorse utilizzate per la sua progettazione. In ambito Ref, la qualità della ricerca viene presentata come l'insieme di tre dimensioni: la qualità dei prodotti (*outputs*), degli effetti (*impact*) e dell'ambiente di ricerca (*environment*); ognuna di esse viene sottoposta a valutazione secondo determinati criteri, definiti da quattro panel disciplinari distinti (A, B, C e D). Per gli interessi relativi al presente testo, è rilevante il lavoro svolto dal panel C (che comprende il settore disciplinare sociologico) sui criteri per la valutazione degli *outputs* (cfr. Ref, 2019). Per fornire un giudizio sulla qualità di una pubblicazione sono stati individuati tre criteri generali, interpretati e chiarificati dal panel in relazione alle specificità dell'area disciplinare di competenza; essi fanno riferimento all'*originalità*, alla *rilevanza* e al *rigore* dei prodotti scientifici e sono stati definiti come segue (*ibidem*):

- *originalità*, intesa come il carattere innovativo della ricerca. Le caratteristiche che contraddistinguono un prodotto originale sono: l'aver affrontato problemi nuovi e/o complessi; lo sviluppo di strategie di ricerca, metodologie e tecniche di analisi innovative; l'acquisizione di nuovo materiale empirico e/o l'avanzamento teorico o l'analisi di una dottrina, delle sue norme o pratiche;
- *rilevanza*, intesa come lo sviluppo di obiettivi intellettuali del campo disciplinare, che può essere teorico, metodologico e/o sostantivo. Il peso relativo sarà assegnato alla rilevanza potenziale, quanto a quella attuale, specialmente se il prodotto è molto recente;
- *rigore*, inteso in termini di precisione intellettuale, robustezza e appropriatezza dei concetti, analisi, teorie e metodologie impiegate nell'ambito della ricerca. Si terrà conto di qualità come l'integrità, la coerenza e la consistenza delle argomentazioni e delle analisi, così come della dovuta considerazione delle questioni etiche.

Affiancando tali definizioni alle dimensioni della qualità della ricerca individuate per la Vqr, sembra potersi cogliere una certa ispirazione italiana al modello anglosassone. In entrambi i casi si fa riferimento all'*originalità* e al *rigore*, ma anche la definizione di *rilevanza*

[5] Il primo esercizio di valutazione anglosassone (con il nome Research Assessment Exercise) risale infatti agli anni '80 (cfr. Bence, Oppenheim, 2005).

 ISBN 9788835103219

sembra in parte sovrapponibile alla definizione di *impatto* italiana. Tuttavia, come si può vedere, la definizione Ref risulta meglio delineata. Per quanto riguarda il carattere innovativo di un prodotto scientifico, si fa riferimento a un ampio elenco di aspetti di una ricerca cui può essere attribuita originalità: problemi, strategie di ricerca, metodologie, tecniche di analisi, materiale empirico, senza che siano trascurati gli elementi di carattere teorico-concettuale. Nel contesto italiano, invece, l'originalità di un prodotto scientifico viene declinata in relazione ad elementi della ricerca alquanto generali, ognuno dei quali richiederebbe di essere specificato: modo di pensare, prospettiva, tesi, tema, fonte, il tutto in relazione a lavori precedenti in rapporto di contiguità con il prodotto valutato. La differenza principale tra le due definizioni, pertanto, risiede nel livello di precisione terminologica-concettuale, indubbiamente maggiore nel contesto anglosassone, grazie all'utilizzo di soluzioni probabilmente più adeguate e in linea con un linguaggio a carattere specialistico.

Anche il concetto di rigore espresso nell'ambito del Ref presenta una maggiore articolazione concettuale, con riferimento a dimensioni quali precisione, robustezza e coerenza, attribuite sia alle modalità di svolgimento teoretico-concettuale del lavoro sia alle procedure della ricerca. Si tratta, quindi, di una definizione che tende ad approfondire, così chiarendo il significato del concetto di rigore. Nel caso italiano, invece, viene utilizzata una strategia alquanto singolare. Innanzitutto, la presenza dell'aggettivo "metodologico" nella denominazione della dimensione sembrerebbe operare una utile delimitazione concettuale, con indubbi benefici in termini di intensione del concetto generale. Nondimeno, la nozione di metodologia, attraverso la quale si ottiene tale riduzione, avrebbe necessitato di una descrizione più accurata. Infatti, da un lato, il rigore metodologico viene declinato nei termini assai generali della capacità di definire chiaramente e raggiungere gli obiettivi della ricerca e del livello di padronanza dello stato dell'arte. Ma lo stesso rigore metodologico viene, d'altra parte, reso con la nozione di "appropriatezza della metodologia", producendosi così un evidente effetto tautologico. Insomma, diversamente dalla definizione anglosassone, si evita accuratamente di indicare quali aspetti metodologici di un lavoro scientifico devono rispondere al requisito del rigore perché esso possa essere giudicato di qualità.

Infine, è possibile affiancare alla dimensione della rilevanza del Ref

la dimensione dell'impatto della Vqr, posto che in entrambi i casi si fa riferimento all'influenza potenziale e attuale che il prodotto scientifico ha esercitato o potrebbe esercitare, la quale nel caso anglosassone può ricadere sul campo disciplinare di pertinenza e, analogamente, nel caso italiano sulla comunità scientifica di riferimento. Tuttavia, le due definizioni differiscono nel momento in cui il Ref precisa che le ricadute di un prodotto rilevante possono essere teoriche, metodologiche e/o sostantive e sono in relazione alla capacità di sviluppare degli obiettivi nel campo disciplinare di pertinenza.

Passando ora alla tematizzazione a scopi valutativi della qualità della ricerca in Francia, in un documento del 2014 (cfr. Hceres, 2014), il Hceres (*Haut conseil de l'évaluation de la recherche et de l'enseignement supérieur*), l'agenzia che ha ufficialmente sostituito l'Aeres, fornisce criteri che tengono conto dell'attività di ricerca anche declinata secondo la struttura organizzativa e la specificità dell'Ente: qualità della produzione scientifica, reputazione e riconoscimento accademico, rapporti con il contesto sociale, economico e culturale, organizzazione e gestione dell'unità di ricerca, sinergia tra attività di ricerca e formazione, strategie di ricerca e prospettive future per i successivi cinque anni.

È interessante notare che il documento è accompagnato da un'avvertenza riguardante il carattere di flessibilità delle procedure indicate; gli standard del Hceres, infatti, non sono imposti nei termini di una rigida griglia di valutazione da seguire pedissequamente, ma come linee guida illustrative, non esaustive, applicabili ad una larga varietà di discipline, che necessitano, in quanto tali, di essere adattate alle caratteristiche specifiche di ogni campo disciplinare (cfr. Hceres, 2014, pp. 5-6). Inoltre, viene dedicato molto spazio alla valutazione della qualità dei prodotti scientifici afferenti alle scienze umane e sociali, considerate un ambito peculiare che risponde a logiche dissimili da quelle della *hard sciences* e che necessita perciò di un'attenzione particolare. Consapevole del tratto distintivo di tali aree disciplinari, il Hceres si è mosso lungo due strade principali, al fine di progettare la strategia valutativa più adeguata. Da un lato, la determinazione delle specificità disciplinari è stata affidata ad una commissione di esperti, definiti "pari", provenienti dalla stessa comunità scientifica delle istituzioni da valutare, in pratica il corrispettivo dei nostri Gev; dall'altro, il tema

dell'adattamento dei criteri generali, comuni di valutazione alle specificità disciplinari è stato affrontato in una serie di discussioni avvenute tra i direttori scientifici del Hceres ed esperti esterni. Il risultato di tale negoziazione non coincide con un elenco di criteri distinti da quelli generali, ma con una serie di standard congiunti, adattabili quando necessario e applicabili anche tenendo conto di istanze, prospettive e obiettivi propri delle scienze umane e sociali (cfr. *ivi.*, pp. 19-20).

Focalizzando l'attenzione sulla valutazione dei prodotti scientifici, il Hceres individua alcune dimensioni della qualità, oggetto di valutazione:

- l'originalità e la portata/rilevanza della ricerca, l'importanza delle scoperte nell'ambito di studi pertinente;
- la realizzazione di importanti scoperte/svolte teoriche e metodologiche, i cambiamenti di paradigma, l'emersione di nuovi problemi o di nuovi percorsi di indagine;
- l'impatto scientifico nell'accademia (citazioni, riferimenti, ecc.);
- i riconoscimenti nazionali o internazionali;
- la reputazione e la selettività della rivista.

Per quanto concerne quest'ultimo punto, si è già detto che l'agenzia francese ha effettuato una classificazione delle riviste, distinguendole in tre classi di merito (cfr. capitolo 1). A tal fine, sempre considerando le specificità del campo disciplinare, sono stati adottati alcuni parametri, non necessariamente esaustivi e allo stesso modo rilevanti per tutti i settori delle scienze umane e sociali. Oltre ai dati anagrafici della rivista (come l'ISSN, il titolo, l'area disciplinare di pertinenza, ecc.), sono state considerate informazioni circa la capacità di diffusione, tra cui la lingua, la regolarità di pubblicazione, la possibilità di accedere alla rivista online, ecc.; informazioni circa la selezione degli articoli, come il tipo di *peer review*, il numero di articoli rifiutati, la trasparenza dei criteri, ecc.; informazioni circa la qualità scientifica, intesa come la presenza di un comitato scientifico e/o editoriale, il tipo di articoli che si è più propensi a pubblicare (rassegne della letteratura, ricerche empiriche, recensioni, ecc.), l'ampio utilizzo di fonti bibliografiche, ecc.; la politica editoriale e la reputazione della rivista.

Tenendo comunque conto che l'esercizio di valutazione francese si

snoda su strade distinte rispetto alla Vqr e ha obiettivi diversi[6], l'utilizzo della classificazione delle riviste scientifiche a fini valutativi rappresenta senza dubbio un elemento che differenzia le due esperienze. Nel caso francese, infatti, viene dichiaratamente realizzata al fine di facilitare la valutazione e l'auto-valutazione, come supporto concettuale agli standard generali individuati. Nel caso italiano, invece, le informazioni circa le riviste vengono relegate tra i metadati di cui possono tenere conto i *referees* nella valutazione. Le riviste scientifiche e di fascia A sono oggi utilizzate ai fini del conteggio delle pubblicazioni relative agli indicatori di impatto dei candidati e dei commissari per l'Asn, ma nulla vieterebbe una loro maggiore valorizzazione da parte dei *referees* nella valutazione dei prodotti conferiti alla Vqr[7], anche considerando l'impegno profuso proprio nella procedura di classificazione[8].

2.3. Un tentativo di chiarificazione concettuale

Alla luce delle considerazioni fin qui esposte, il processo di chiarificazione del concetto di qualità di un prodotto scientifico assume il ruolo probabilmente più importante dell'intero esercizio di valutazione, poiché dal suo esito dipenderanno tutte le fasi successive. Il processo di chiarificazione concettuale, inoltre, rappresenta un'occasione per le comunità scientifiche afferenti alle diverse Aree disciplinari di partecipare al processo di valutazione che le interessa e di mettere a punto una definizione condivisa del concetto e dei criteri attraverso i quali valutare la qualità dei prodotti scientifici.

Si è già visto come nella formulazione del bando Vqr e nelle linee guida l'Anvur abbia fornito una definizione del concetto di qualità dei prodotti scientifici di natura stipulativa, focalizzata prevalentemente su tre criteri, i cui confini tuttavia restano vaghi. È altresì doveroso aggiungere che la definizione dell'Anvur doveva essere allo stesso

[6] Lo scopo dell'agenzia francese sembra essere primariamente legato a un programma in grado di valorizzare pratiche di autovalutazione da parte di enti di ricerca accademici, allargando lo sguardo ad altre dimensioni della ricerca scientifica che non siano rappresentate dalla pubblicazione dei suoi risultati (cfr. Hceres, 2014, p. 4).

[7] Si potrebbe immaginare anche un ruolo di mero supporto nella risoluzione dei casi di discordia tra i giudizi dei due *referees*.

[8] Si pensi che le riviste presenti negli elenchi di fascia A superano le 6000 unità, mentre le riviste considerate scientifiche sono oltre 25000.

tempo generale ma non generica, tanto da poter essere applicata alle 16 Aree ministeriali (cfr. Fasanella, Martire, 2017, p. 90). A fronte dell'evidente carenza definitoria del bando e delle linee guida rispetto ai tre criteri costitutivi della qualità di un prodotto scientifico (cfr. Tab. 1), alcuni Gev hanno condotto uno sforzo di esplicitazione di tale concetto con riferimento al proprio settore disciplinare, come nel caso del Gev di Area 12, quella giuridica (cfr. Anvur, 2017l; v. anche capitolo 5), che è così giunto ad una definizione di qualità più chiara e al tempo stesso più specifica rispetto ad altre Aree e in particolare rispetto all'Area 14, delle scienze politiche e sociali.

Il percorso di chiarificazione del concetto di qualità dei prodotti scientifici che verrà qui proposto, seppure riferito specificamente all'Area 14, non si esclude possa fornire spunti di riflessione per altre Aree. Conseguentemente a quanto più sopra discusso, si punterà all'individuazione di dimensioni concettuali più specifiche che possano favorire l'avvicinamento al piano dell'osservazione e perciò idealmente facilitare il processo della valutazione e renderlo più difendibile sotto l'aspetto della confrontabilità inter ed intra soggettiva (cfr. più sopra). In questo percorso, inoltre, saranno agevolmente riconoscibili suggestioni tratte dalle esperienze di valutazione della ricerca inglese e francese, che, come illustrato nelle pagine precedenti, costituiscono i riferimenti principali per lo sviluppo della Vqr italiana.

Nell'area delle scienze politiche e sociali il concetto di qualità di un prodotto scientifico potrebbe essere reso attraverso alcune dimensioni fondamentali: *originalità*; *accuratezza metodologica*; *appropriatezza della scrittura*; *riferimento alle fonti*; *adeguatezza della trattazione*; *reputazione e selettività della rivista*. Tuttavia, l'operativizzazione delle dimensioni individuate impone una necessaria, preliminare articolazione concettuale di ciascuna di esse in componenti semanticamente congruenti.

L'originalità di un prodotto può essere riferita al piano empirico, a quello teorico-interpretativo, a quello della tematica trattata e infine al piano metodologico. L'originalità empirica può essere valutata rispetto a un prodotto scientifico che fornisca un contributo sul piano dell'applicazione di una o più teorie consolidate a nuove evidenze empiriche. L'originalità teorico-interpretativa ha a che vedere con la capacità di un prodotto di fornire un contributo attraverso la formulazione di concetti e/o di ipotesi e/o di teorie inediti con riferimento al tema

trattato. L'originalità tematica riguarda l'attitudine di un prodotto a trattare nuovi temi o temi di recente introduzione nel dibattito scientifico. Infine, l'originalità metodologica riguarda la presenza nell'ambito del prodotto da valutare di soluzioni procedurali innovative sul versante della progettazione del disegno di ricerca e/o sul versante delle tecniche di raccolta dei dati e/o sul versante delle tecniche di elaborazione e analisi dei dati.

Un prodotto scientifico di qualità sotto l'aspetto dell'accuratezza metodologica sarà contraddistinto da un'impostazione del tema/problema oggetto dell'indagine e/o della trattazione che risulti logicamente corretta e consequenziale. Il disegno della ricerca/della trattazione dovrà risultare adeguato rispetto alla domanda cognitiva che ne sta alla base; le tecniche di raccolta dei dati, di elaborazione e analisi dovranno essere utilizzate correttamente.

L'appropriatezza della scrittura, che rappresenta una dimensione inedita rispetto alla definizione adottata dall'Anvur, si può intendere, per un verso, come qualità redazionale complessiva del testo, a coprire sia il piano grammaticale sia l'aspetto della linearità e della scorrevolezza; per altro verso, può essere valutato il ricorso nella redazione del testo non solo a concetti ma anche a termini specialistici, familiari al campo disciplinare di riferimento.

Un'ulteriore dimensione concettuale che non era stata presa in considerazione in maniera esplicita è rappresentata dal riferimento alle fonti, in una duplice accezione. Si può valutare, in primo luogo, se la trattazione del problema/tema affrontato sia basata su adeguati riferimenti alla letteratura specialistica del settore; in secondo luogo, se i riferimenti alla letteratura e ad altre possibili fonti ne consentono una ricostruzione il più possibile puntuale.

L'adeguatezza della trattazione potrebbe essere resa semanticamente facendo riferimento a tre diverse componenti. Anzitutto, si potrebbe valutare se le argomentazioni addotte a sostegno delle posizioni presentate nel corso della trattazione siano collegate a qualche evidenza diretta o indiretta a cui la trattazione stessa rinvia. In secondo luogo, si potrebbe stabilire se le argomentazioni addotte a sostegno delle posizioni presentate nel corso della trattazione siano realmente collegate alle fonti bibliografiche citate. Il terzo aspetto oggetto di valutazione potrebbe rinviare alla sostenibilità, intesa come la coerenza sul piano meramente logico delle posizioni presentate.

 ISBN 9788835103219

Infine, limitatamente agli articoli pubblicati su rivista, un possibile elemento da considerare in vista del perfezionamento del processo di valutazione potrebbe essere la collocazione delle riviste secondo la classificazione operata da Anvur sulla base della reputazione e della selettività della rivista stessa, così come avviene per il modello francese discusso nelle pagine precedenti. Naturalmente, una decisione di tal genere presuppone un'attenta, trasparente, condivisa attività di monitoraggio e valutazione del lavoro delle riviste stesse, con lo scopo di aggiornare continuamente la classificazione già utilizzata per l'Asn (dal lato sia dei candidati sia dei commissari) e da utilizzare eventualmente per la Vqr. In effetti, un lavoro di tal genere è formalmente già svolto dall'agenzia di valutazione e si avvale di gruppi di esperti per ciascuna area disciplinare. Tali esperti hanno proprio il compito di assegnare la rivista a una delle due classi identificate (scientifica e di classe A) secondo criteri determinati, procedendo ciclicamente ad un aggiornamento, sia valutando richieste di afferenza ad una classe, sia verificando la sussistenza delle condizioni di permanenza delle riviste già posizionate in una delle suddette classi. Vale la pena di sottolineare che una soluzione di questo tipo realizzerebbe una rivoluzione del modello attualmente adottato. Mentre ora, discutibilmente, il punteggio Vqr dei prodotti pubblicati in una data rivista viene utilizzato come uno dei criteri per l'assegnazione o la conferma della rivista nella classe A, nella prospettiva qui delineata la pubblicazione di un dato prodotto da parte di una rivista di classe A varrebbe come uno dei parametri che contribuiscono alla determinazione del punteggio Vqr di quello stesso prodotto.

Una svolta del genere avrebbe senso se solo si consideri l'attitudine delle riviste, in special modo le riviste scientifiche e di classe A, a restituire la qualità reale della ricerca, ovvero il significato applicato di tale nozione, nelle sue molteplici e talora molto differenziate sfaccettature, anche rispetto a uno stesso, specifico settore disciplinare. Ovviamente un'opzione di tal genere impegna le riviste a rendere massimante trasparenti e *accountable* tutti i passaggi compresi tra la *submission* di una proposta editoriale e la sua pubblicazione, o mancata pubblicazione.

Nella proposta presentata, come si può vedere, non figurano riferimenti all'impatto del prodotto sulla comunità scientifica di settore. La ragione di tale omissione è fondamentalmente logica. L'impatto, infatti, è una conseguenza della qualità ma non può essere considerato

nei termini di una dimensione costitutiva di essa, sicché si può affermare che la qualità implica l'impatto. Applicando semplicemente un ragionamento in termini di *modus ponens*, una volta accertata la qualità di un prodotto non avrebbe senso perciò valutarne l'impatto. Certo, si potrebbe ragionare in termini di *modus tollens*, e cioè, valutando l'assenza di impatto di un prodotto, si potrebbe concludere che non si tratta di un prodotto di qualità. Rimarrebbero però aperte due questioni. La prima: come valutare in termini di impatto un prodotto che non sia di qualità? La seconda: come valutare in termini di qualità un prodotto che abbia impatto? Ovviamente la soluzione di tali questioni, sul piano strettamente logico, consisterebbe nello stabilire una relazione di doppia implicazione tra qualità e impatto: *se e solo se* un prodotto è di qualità allora ha (avrà) impatto. Una condizione di questo genere risulta, tuttavia, manifestamente insostenibile.

In verità, c'è da ritenere che la relazione tra le due dimensioni, qualità e impatto, sia di natura probabilistica. La qualità di un prodotto è uno dei fattori, peraltro non necessario, dell'impatto, ma evidentemente non è l'unico. A tale proposito, basti qui ricordare il dibattito sul carattere spurio delle citazioni quali elementi per lo stabilimento, per via abduttiva, del valore di un prodotto (Bornmann, Daniel 2008; Costas, Bordons, van Leeuwen, van Raan, 2009; Glänzel, 2008; Hicks, Wouters, Waltman, de Rijcke, Rafols, 2015; Merton, 1968; Penfield, Baker, Scoble, Wykes, 2014; Seglen, 1997; Aksnes, Langfeldt, Wouters, 2019).

Per concludere, vale la pena soffermarsi su un punto che in questa sede potrà essere solo sollevato, rinviando a un'altra occasione una riflessione più articolata e senza dubbio impegnativa. Le componenti della qualità scientifica qui presentate possono verosimilmente essere sottoposte a un processo di ponderazione. In altre parole, il peso del contributo fornito da ciascuna componente alla determinazione della qualità scientifica di una pubblicazione potrebbe non essere lo stesso. Inoltre, questa stessa logica potrebbe essere applicata a ciascuna delle subcomponenti in cui si articola, come abbiamo visto, ognuna delle componenti principali del concetto di qualità scientifica. Seppure al concetto generale di qualità scientifica possa essere attribuita una certa stabilità intersettoriale e intrasettoriale (ma anche di questo varrebbe la pena discutere!), un po' meno agevolmente è possibile affermare

 ISBN 9788835103219

tale principio di stabilità con riferimento al peso delle relative dimensioni e subdimensioni. Non si può escludere, infatti, che il contributo ponderato che, poniamo, l'*originalità* fornisce al concetto di qualità scientifica possa variare da un settore scientifico-disciplinare a un altro. Per esempio, in un settore di recente costituzione, l'*originalità* potrebbe ai fini della qualità essere meno rilevante dell'*accuratezza metodologica*, mentre il discorso inverso potrebbe valere per un settore altamente consolidato. Tale variabilità potrebbe registrarsi anche nell'ambito dello stesso settore in momenti storici diversi, senza contare che, pure in una dimensione di sincronismo, entro un medesimo settore potrebbero sussistere campi di interesse scientifico più ristretti e non perfettamente uniformati alla definizione di qualità corrente al livello più alto. Naturalmente, questa stessa logica investe anche le subdimensioni di ciascuno degli aspetti della qualità considerati. Per come esse sono state più sopra definite, si può immaginare, per esempio, che, anche entro uno stesso settore scientifico-disciplinare, l'*originalità empirica* possa rivestire un'importanza cruciale nei periodi di scienza normale, a scapito dell'*originalità teorico-interpretativa*; verosimilmente, il rapporto tra i due aspetti si invertirebbe nei periodi di scienza rivoluzionaria, nei quali, invece, la capacità di fornire risposte teoriche innovative e risolutive a problemi aperti e non chiusi entro i quadri di senso correnti risulterebbe molto più apprezzata (Kuhn, 1962; tr. it., 1978).

Queste ultime considerazioni riportano alle preoccupazioni avanzate in apertura di capitolo. Il concetto di qualità della ricerca, come si è visto, si presenta astratto, complesso, multiponderale. L'applicazione controllata di esso ai fini della valutazione richiede un processo lungo, impegnativo, a carattere negoziale e niente affatto scontato di lenta metabolizzazione, che non escluda alcuno degli attori coinvolti (autori, revisori, esperti valutatori, decisori). Solo un processo di questo genere, infatti, può favorire la condivisione di una data, qualsivoglia, nozione di qualità della ricerca, non solo sul piano generale, ma in tutte le sue numerose e diversamente incidenti sfaccettature, che, come si è detto, determinano differenze sia tra i diversi settori scientifico-disciplinari sia all'interno dello stesso settore. Se si rinunciasse a lavorare in questa faticosa direzione di valutazione partecipata (cfr. Palumbo, 2003; Palumbo, Torrigiani, 2009) per cedere a soluzioni

solo in apparenza semplificatorie, per quanto di sicuro più veloci, basate su un assunto non fondato di *tacit knowledge* (Polanyi, 1966; tr. it., 1979), si rischierebbe di delegittimare la valutazione. Essa risulterebbe così liberata da vincoli di controllo e replicabilità ma inevitabilmente sospinta verso pratiche che valorizzano una certa, malintesa soggettività, e determinano, di fatto, la dismissione del principio di determinatezza e uniformità d'uso, che rappresenta indiscutibilmente il prerequisito di qualunque, riconoscibile processo di valutazione.

 ISBN 9788835103219

3. I Gev e i revisori

di Lorenzo Barbanera

Qualsiasi gruppo di individui che si determina in un'organizzazione sociale è chiamato a stabilire un punto d'equilibrio ottimale tra la libertà d'azione del singolo e i vincoli normativi – taciti o scritti – che limitano quella stessa libertà al fine di garantire la sopravvivenza e il normale funzionamento della società. Pertanto, il sistema Vqr e tutti i suoi attori non possono esimersi dall'intraprendere un percorso riflessivo che abbia come obiettivo quello di garantire la *giusta distanza* dai due poli. Il primo designa uno scenario dove l'assenza di un nucleo fondante di procedure chiare e condivise genera margini di discrezionalità incontrollati, mettendo così a repentaglio la comparabilità dei giudizi nonché la capacità di rendicontare il lavoro svolto e assicurare gli opportuni standard in termini di *accountability*. D'altro canto, un'eccessiva irreggimentazione sostenuta da un'acritica applicazione di procedure minutamente definite rischia di mortificare l'opera di coloro che si cimentano nella valutazione della qualità della ricerca. Infatti, l'esercizio valutativo richiede ai revisori di mettere in gioco risorse e abilità parzialmente legate a una conoscenza tacita sedimentata nell'esperienza che, in virtù di questo suo radicamento, è inestricabilmente connessa alla loro soggettività. Benché si tratti di un fattore sfuggente, implicito e dunque arduo da controllare e regolare, metterlo al bando significherebbe riproporre un'idea di baconiana memoria per cui chiunque potrebbe svolgere il lavoro dell'esperto valutatore, a patto di seguire in modo pedissequo una metodologia prescrittiva rigidamente e minuziosamente strutturata. D'altronde, sarebbe pericoloso, oltre che ingenuo, credere di poter individuare il perfetto procedimento che conduce alla perfetta rilevazione della qualità, concetto per sua

 ISBN 9788835103219

natura etereo e dunque forzatamente negoziale. (Di Benedetto, 2015, p. 96).

Uno dei presupposti, dunque, è il pieno riconoscimento delle *variabili umane*; ossia quei fattori di soggettività che inevitabilmente permeano sia la ricerca scientifica che quella valutativa, per quanto entrambe tendano all'oggettività attraverso l'esercizio del pensiero razionale e del rigore procedurale. Peraltro, un simile presupposto può essere attualizzato con sufficiente agilità nel dominio delle scienze sociali e, in particolare, nell'ambito della sociologia. Infatti, l'immaginazione sociologica, che secondo Mills «permette a chi la possiede di vedere e valutare il grande contesto dei fatti storici nei suoi riflessi sulla vita interiore e sul comportamento esteriore di tutta una serie di categorie umane» (Wright Mills, 1959; tr. it., 1962, p. 15), si configura come un atto creativo ed esclusivo, nel senso che, come dice Wright Mills, non tutti possiedono l'immaginazione sociologica. Viceversa, la pratica scientifica di baconiana memoria, fondata su un dispositivo normativo ineluttabile e sempre uguale a se stesso, è del tutto inclusiva in quanto non richiede immaginazione, bensì meri esecutori; individui a cui viene richiesto di mettere in pratica un *corpus* predeterminato di regole al fine di conseguire, tutti indistintamente, lo stesso *identico* risultato.

Nondimeno, è alquanto improbabile che un simile principio possa realizzarsi, poiché, malgrado gli opportuni sforzi vòlti al raggiungimento di un sapere stabile e oggettivo, da sempre l'irriducibile pluralità dei punti di vista ha generato diverse concezioni filosofiche del mondo, da cui sostanziano altrettante tradizioni sociologiche (cfr. Collins, 1994). In effetti, sin dagli albori si assiste al confronto tra paradigmi, approcci e scuole, un dibattito che talvolta assume le sembianze di un vero e proprio scontro frontale; si veda, ad esempio, l'eterna e per molti versi infruttuosa diatriba fra qualità e quantità.

Per questo motivo la Vqr deve affrontare, volente o nolente, la frammentarietà dovuta alle molteplici contrapposizioni disciplinari tipiche dell'Area 14, soprattutto con riferimento al settore sociologico che, citando l'Anvur, si caratterizza per «una notevole eterogeneità, sia tra i diversi Ssd (Settori Scientifico Disciplinari) che, spesso, dentro ciascun Ssd. Eterogeneità per dimensioni quantitative, per contenuti tematici, per orientamenti metodologici, per strategie di pubblicazione» (Anvur, 2017a, p. 21). Smarrire la consapevolezza di questa

eterogeneità provocherebbe seri danni, poiché è necessario far sì che essa rientri a pieno titolo tra i fattori che determinano il processo di valutazione della ricerca in tutte le sue fasi costitutive.

Pertanto, pur sapendo che «la valutazione *peer* potrebbe essere *biased*, soprattutto in settori metodologicamente divisi in scuole, e privi di standard comuni legittimati dall'insieme degli addetti» (*ivi*, p. 42), sarebbe quantomeno azzardato pensare di risolvere la questione proponendo protocolli che mortificano il potenziale arricchente delle variabili umane in gioco, sacrificandole sull'altare di procedure valutative che mirano alla standardizzazione completa promettendo una maggiore oggettività dei risultati. In altri termini, la soluzione uniformante, in uno scenario per sua natura multiforme, è destinata allo scacco.

Di conseguenza, non convince appieno il ricorso alla bibliometria, strumento dimostratosi largamente inefficace se applicato alle scienze sociali (cfr. Garfield, 1979; Campelli, 2011), sia per motivi legati alle differenti forme d'uso della citazione, sia perché, in ogni caso, «le citazioni, anche quelle «autentiche» non possono essere semplicemente contate: vanno lette e *valutate*» (Campelli, 2011, p. 7). E così, quando si lascia intendere che le distorsioni a carico della *peer review* potrebbero essere contrastate «disponendo di criteri oggettivi, quali quelli bibliometrici» (Anvur, 2017a, p. 42), permane un certo scetticismo. In sostanza, se da un lato l'idea di sfruttare la bibliometria «per avere un termine di confronto da porre in equilibrio riflessivo con le valutazioni *peer*» (*ibid*.) risulta senz'altro bilanciata e prudente, dall'altro il suo potenziale euristico rimane, comunque, limitato[1].

Dunque occorre affrancarsi dall'ossessione di eliminare ogni fattore di soggettività per concentrare gli sforzi verso soluzioni che prevedano, al contrario, di *accogliere* quella stessa soggettività, attribuendole un ruolo che la valorizzi e al contempo ne circoscriva, per quanto possibile, gli effetti indesiderati.

In tal senso, si potrebbe migliorare la redazione delle linee guida per renderle realmente integrative rispetto alle indicazioni del Bando, evitando che con esso intrattengano un rapporto di sostanziale circolarità, come invece è accaduto nell'ultimo esercizio valutativo. Per

[1] Per ulteriori approfondimenti sul tema della bibliometria si rimanda al capitolo 1 del presente lavoro.

esempio, sarebbe appropriato mettere a punto delle linee guida all'uso pratico del concetto di qualità della ricerca, in assenza delle quali lo sforzo definitorio iniziale rischia di non risultare effettivo; così come si potrebbe pensare a delle linee guida per la stesura di un commento a latere che giustifichi il punteggio espresso per ognuna delle dimensioni che compongono il concetto.

Applicando i principi appena illustrati, gli esperti valutatori e i revisori verrebbero messi nella condizione di svolgere al meglio il loro compito senza correre il rischio di disattendere istanze irrealistiche come, per esempio, mettersi *idealmente* «*sotto un velo di ignoranza*, evitando di farsi influenzare da fattori relativi alla persona che ha prodotto il lavoro» (Anvur, 2017b, p. 11). Qui si apre un argomento delicato, che chiama direttamente in causa il problema irrisolto del conflitto fra scuole e del suo riflettersi nella definizione del concetto di "parità". Rinviando al lavoro di Di Benedetto (2015, p. 48) per ulteriori approfondimenti sul tema, in questa sede si vuole solo mettere in evidenza la necessità di rinnovare lo sforzo riflessivo sulla nozione di parità, riconsiderando il peso che l'appartenenza o meno dei revisori a micro-comunità, correnti e *background* diversi può esercitare sugli esiti della *peer review*. In realtà, trovare il giusto equilibrio è molto complesso: in un caso, infatti, perseverare nell'idea di un essenziale isomorfismo tra la comunità scientifica e il gruppo dei pari significherebbe esporsi alle distorsioni dovute alle numerose incompatibilità – talvolta alle incomprensioni – fra le varie scuole; nell'altro, l'eccessiva restrizione del dominio della comunità scientifica aumenterebbe l'omogeneità interna al gruppo, rischiando di ridurre i giudizi a un concerto di voci assenzienti.

In conclusione, avanzando sul solco di queste brevi considerazioni, il presente lavoro si impernia su una logica della *giusta distanza*, volendo evitare sia l'indiscriminata meccanizzazione della Vqr, sia la sua totale sottomissione alle idiosincrasie individuali degli attori coinvolti. Tale volontà muove da una prospettiva orientata positivamente che afferma la necessità di sottoporre a valutazione gli esiti della ricerca accademica e, al contempo, ribadisce l'esigenza di evidenziare i punti deboli delle procedure valutative al fine di favorirne il costante miglioramento.

 ISBN 9788835103219

3.1. I Gruppi di Esperti Valutatori

3.1.1. La composizione dei gruppi

Come disponeva il Bando di partecipazione, gli esperti valutatori (d'ora in avanti Ev) sono stati nominati dal consiglio direttivo dell'Anvur che ha selezionato gli «studiosi di elevata qualificazione scelti sulla base dell'esperienza internazionale nel campo della ricerca e alle esperienze di valutazione già compiute» (Anvur, 2015, p. 2). Opportunamente, l'Anvur ha integrato questa descrizione sommaria con un documento che specificava le procedure utilizzate per la selezione dei componenti del Gev. Nel dettaglio, si fa riferimento a tre caratteristiche che hanno concorso alla scelta degli esperti:

1. qualità scientifica (misurata, ove possibile, tramite l'*h-index*, il numero totale di citazioni, eventuali riconoscimenti del merito scientifico, analisi degli elementi di curriculum vitae presenti nella manifestazione di interesse, ecc.);
2. continuità della produzione scientifica negli ultimi cinque anni;
3. esperienza in attività di valutazione a livello nazionale o internazionale (Anvur, 2015a, p. 14).

La specificazione dei requisiti si pone così ad un buon livello di chiarezza; nondimeno, vi sono margini per un ulteriore sforzo definitorio. Ad esempio, si potrebbe esplicitare se la "*continuità* della produzione scientifica" sia intesa in un'ottica puramente quantitativa legata alla *produttività* o in un senso più qualitativo di *coerenza* rispetto a un certo ambito di ricerca, oppure si componga di un mix di entrambe le accezioni. Inoltre, soprattutto in assenza dell'*h-index*, sarebbe bene indicare che tipo di "riconoscimenti del merito scientifico" e di "elementi del curriculum" favoriscono l'accesso al ruolo di Ev. Infine, si potrebbe indicare nel Bando quali "esperienze in attività di valutazione" rientrano nel computo di quelle valide ai fini della selezione.

La specificazione delle caratteristiche che i candidati a esperto valutatore devono avere potrebbe condurre alla formulazione di requisiti più peculiari inseriti nel contesto di un sistema analogo a quello dell'Asn (Abilitazione Scientifica Nazionale), attraverso cui l'Anvur, di

concerto con le comunità scientifiche, arriverebbe a definire un insieme di criteri pubblici con lo scopo di fissare delle soglie minime per l'accesso al ruolo di esperto. Infine, tutti gli idonei potrebbero essere inseriti in una lista da cui attingere per dare forma ai Gruppi di Esperti della Valutazione relativi ad ogni area Cun.

Posto che la scelta di rendersi o meno eleggibili rimarrebbe in ogni caso un appannaggio della volontà del singolo, questa soluzione ovviamente non risolverebbe l'inevitabile elemento distorsivo connesso all'autoselezione dei candidati (cfr. Fasanella e Di Benedetto, 2015, p. 47). Ciò nonostante, offrirebbe maggiori garanzie in termini di *accountability*, soprattutto laddove si accogliesse la proposta di pubblicare l'elenco dei revisori selezionati e di coloro che non hanno ottenuto l'idoneità.

Per quanto concerne la numerosità dei Gev, essa è stata fissata «sulla base del numero atteso di prodotti da valutare nelle diverse aree e della percentuale prevista di prodotti da sottoporre a *peer review*» (i*vi*, p. 3). Tuttavia, in questo modo non si hanno informazioni sufficienti né da un punto di vista quantitativo, in quanto vi è sempre uno scarto fra prodotti attesi e conferiti, né in un'ottica qualitativa, perché rimane ignoto il contenuto degli stessi.

In realtà, per quanto riguarda la differenza numerica tra prodotti attesi e conferiti non sembrano esserci misure risolutive, in quanto il divario può essere dovuto a molteplici fattori – spesso di natura contingente – del tutto imprevedibili. Una soluzione potrebbe consistere nel posticipare la formazione dei Gev in un momento successivo all'acquisizione dei prodotti, ma questo rischierebbe di allungare le tempistiche dell'intero esercizio valutativo, comprimendo, al contempo, quelle concesse ai *referees* per effettuare le revisioni.

Nondimeno, sembra possibile intervenire sul versante qualitativo del problema. Nello specifico, si potrebbero sfruttare le informazioni relative ai Ssd dei prodotti attesi, in modo da comporre i Gev cercando di mantenere un certo grado di proporzionalità tra l'afferenza disciplinare dei contributi e le competenze degli Ev. Questo permetterebbe da un lato, di arginare possibili carenze di figure competenti in specifici ambiti disciplinari, dall'altro, reciprocamente, di evitare inutili eccedenze.

Prima di procedere oltre, è bene ricordare che le misure correttive qui proposte non possono essere considerate di per sé definitive e risolutive. Esse devono inserirsi in un sistema di contromisure organico

e sinergico che tenga conto del più ampio contesto della Vqr, le cui criticità si legano trasversalmente a tutte le fasi che la caratterizzano.

3.1.2. L'importanza dell'eterogeneità dei gruppi

Gli Ev ricoprono un ruolo di assoluta centralità nell'ambito della Vqr. Essi, come previsto dal Bando, si fanno carico di attività che implicano responsabilità cruciali; prime fra tutte l'attribuzione dei prodotti ai *referees* e la gestione delle controversie. Pertanto, in assenza di una procedura di selezione degli Ev rigorosa, condivisa e controllabile, si rischia di introdurre effetti distorsivi capaci di inficiare l'intero processo valutativo. Non a caso, dunque, nel d.m. per le linee guida della valutazione è scritto chiaramente che la presenza di esperti deve essere «adeguata ed equilibrata» (d.m. n. 458 del 2015).

Si osservi ora la tabella che riporta la composizione del Gev per l'Area 14. Anzitutto, si tratta di un insieme di persone numericamente ristretto, e ciò rende ancor più impellente garantire la massima eterogeneità possibile, in quanto, per ovvie ragioni statistiche, vi è un rapporto di proporzionalità inversa tra il numero di membri e l'influenza imprevista di eventuali variabili sovrarappresentate o sottorappresentate. Purtroppo, da questo punto di vista, il Gev 14 sembra essere stato almeno parzialmente deficitario. Nello specifico, il problema principale ha riguardato l'affiliazione degli Ev, alcuni dei quali provenienti dal medesimo ateneo, come accaduto per le università di Trento e Roma Tre, cui corrispondevano due Ev ciascuna (Tab. 1). In aggiunta, quattro esperti[2] avevano maturato una parte importante della loro formazione professionale all'università di Trento, sebbene in quel momento lavorassero in altri atenei[3].

[2] Delia Stefania Baldassari, Peter Wagner, Lorenzo Bordogna e Marco Meriggi.

[3] È opportuno specificare che il problema si reputa tale indipendentemente dalle specifiche università e dai singoli attori coinvolti. Ciò che si vuole far valere, in sostanza, è un concetto generale per cui sarebbe bene che nessun ateneo, a prescindere dal nome, fosse sovrarappresentato.

 ISBN 9788835103219

Tab. 1 - Composizione del Gruppo di Esperti della Valutazione

Cognome	Nome	Affiliazione
Galeotti	Anna Elisabetta	Piemonte Orientale
Andrini	Simona	Roma Tre
Attinà	Fulvio	Catania
Baldassarri	Delia Stefania	New York University
Bordogna	Lorenzo	Uni Milano
Buzzi*	Carlo	Uni Trento
Colombo	Fausto	Uni Cattolica del Sacro Cuore
Meriggi	Marco	Uni Napoli Federico II
Nuti	Leopoldo	Roma Tre
Piattoni	Simona	Uni Trento
Tuccari	Francesco	Uni Torino
Wagner	Peter	Universitat de Barcelona

* Incluso nel Gev da marzo 2016
Fonte: Anvur, 2017a

Un simile sbilanciamento, per quanto lieve a prima vista, può tuttavia aver comportato effetti distorsivi rilevanti. In primo luogo, soprattutto nell'Area 14, contraddistinta dalla ricchezza delle tradizioni scientifiche e culturali presenti nella comunità, è fondamentale che il gruppo di esperti sia massimamente eterogeneo, con l'intento di garantire – o tendere a garantire – il giusto riconoscimento a tutte le principali comunità del settore. Diversamente, laddove si creassero le condizioni per una rappresentatività limitata, si potrebbe favorire la prevalenza di una o più scuole di pensiero, penalizzando ovvero sostenendo determinate classi di prodotti non già per motivi meramente contenutistici, bensì per la loro adesione a una scuola piuttosto che a un'altra.

Si potrebbe ragionevolmente obiettare che si tratta di un traguardo inarrivabile, adducendo come argomento l'impossibilità di offrire spazio a tutte le molteplici prospettive disciplinari per via dell'esiguità numerica del Gev 14 che, nell'ultima Vqr, si componeva di dodici persone. Sarebbe allora auspicabile prevedere la partecipazione di più esperti all'interno del gruppo – e, forse, anche evitare che vi sia più di un componente per ogni ateneo – in modo da consentire una sua composizione ulteriormente ponderata e inclusiva. Senza dubbio, tale in-

 ISBN 9788835103219

clusività non potrà mai essere raggiunta in senso assoluto, ma d'altronde sarebbe ingenuo anche solo pensarlo. Comunque sia, permane l'esigenza di una più equa distribuzione dei membri del Gev.

Vi è poi una seconda ragione per cui è indispensabile garantire l'eterogeneità del Gev. Fra le varie mansioni loro assegnate, gli Ev hanno il compito di definire «di concerto con l'Anvur, i criteri utilizzati per la valutazione dei prodotti» (Anvur 2015, p. 14). Con tutta evidenza, la redazione dei criteri valutativi assume un'importanza capitale nella misura in cui essi costituiscono l'imprescindibile riferimento di ogni revisore, che si impegna a sviluppare il giudizio sulla base della formulazione dei criteri stessi. Dunque, il fatto di declinarli in un modo o nell'altro può incidere profondamente sugli esiti della valutazione. E poiché, verosimilmente, ogni Ev fornisce il suo contributo attingendo al proprio *background* di conoscenze e competenze, è bene far sì che al momento della stipulazione dei criteri possano incontrarsi più punti di vista differenti, a garanzia di una maggiore imparzialità ed esaustività.

In ultima istanza, si vogliono sottolineare i potenziali benefici correlati all'acquisizione di un livello ottimale di rappresentatività. Uno dei principali problemi emersi non solo nell'Area 14, ma in tutte le aree Cun, riguarda il grado di accordo fra i punteggi assegnati ad uno stesso contributo da parte dei due revisori chiamati a valutarlo[4]. Per ora è sufficiente evidenziare che i casi discordanti, cioè quelli in cui le due valutazioni si distanziano per una o più classi di merito (fino a un massimo di quattro), sono stati molto più numerosi di quelli concordanti; nello specifico, solo il 33,7% dei prodotti sottomessi a *peer review* ha ottenuto giudizi unanimi, e ciò significa che circa i due terzi dei prodotti totali sono stati valutati dai *referees* in maniera difforme[5].

[4] Il fatto che le controversie siano numericamente simili in tutte le aree può mettere in discussione l'ipotesi secondo cui, principalmente, esse dipendono dalla spiccata pluralità disciplinare di alcune aree rispetto alle altre. Per mantenere viva l'ipotesi, sarebbe necessario riconoscere a ognuna di esse un determinato grado di frammentarietà interna capace di giustificare le controversie. Tuttavia, si potrebbe anche imputare l'incidenza dei casi discordanti ad una mancata condivisione del concetto di qualità da parte dei *referees*, la cui definizione, ponendosi ad un livello troppo generale, non sarebbe idonea per un'applicazione trasversale a tutte le aree (cfr. capitolo 1). Infine, una quota delle controversie potrebbe essere imputata alla differenza di competenze fra i due *referees* chiamati a giudicare lo stesso contributo. In tal caso, la difformità dei giudizi rifletterebbe non tanto l'appartenenza a diverse scuole di pensiero, quanto l'effettivo grado di prossimità fra le tematiche del prodotto e il background disciplinare dei revisori.

[5] Per una trattazione esaustiva di questa tematica si rimanda al capitolo 5.

Tab. 2 - Numero e val. % di revisioni peer discordanti per 1, 2, 3 e 4 classi per l'area 14

Area 14	Prodotti sottoposti alla *peer review*	Di cui con valutazioni concordanti	Di cui con valutazioni discordanti di 1 classe	Di cui con valutazioni discordanti di 2 classi	Di cui con valutazioni discordanti di 3 classi	Di cui con valutazioni discordanti di 4 classi
Valori assoluti	2.953	995	1.254	532	156	16
% sul totale	100	33,7	42,5	18,0	5,3	0,5

Fonte: Anvur, 2017a

Come si vedrà più avanti, questa situazione può essere considerata sintomatica di un problema più ampio legato alle procedure di *matching* fra prodotto e revisore. Tuttavia, per ora, è sufficiente osservare che una maggiore eterogeneità del Gev avrebbe aiutato, nella fase di gestione e smistamento dei prodotti, a evitare che i membri si fossero trovati nella condizione di «distribuire per la valutazione prodotti di cui non sono personalmente esperti e di cui non conoscono potenziali valutatori competenti sul tema» (Anvur, 2017a, p. 70). Un Gev adeguatamente eterogeneo sul piano delle competenze disciplinari creerebbe i presupposti per una migliore gestione della pluralità intra-disciplinare, riuscendo a limitare situazioni eccessivamente discordanti e conflittuali legate alla valutazione finale dei prodotti.

3.1.3. L'assegnazione dei prodotti agli Ev

Un momento cruciale dell'esercizio valutativo è stato l'attribuzione dei prodotti agli Ev i quali, a loro volta, dovevano assegnarli ai revisori. In realtà, l'iter che ha permesso ai contributi di giungere nelle mani dei *referees* si è composto di varie fasi illustrate come segue: l'iniziale conferimento «dei prodotti di ricerca al Gev si basa sul Ssd dell'addetto»[6] (Anvur, 2017b, p. 7); successivamente, «Il Gev 14 dividerà i prodotti scientifici per tipo di pubblicazione e area di ricerca

[6] Con questo termine si intende «il personale incardinato nelle istituzioni cui sono stati associati i prodotti di ricerca da valutare» (Anvur, 2017a, p. 9). In altri termini, gli autori dei prodotti.

 ISBN 9788835103219

e li assegnerà al Sub-Gev più appropriato» (*ibid.*); a seguire «Il Coordinatore di Sub-Gev[7] li affiderà a due componenti di Sub-Gev sulla base del criterio di maggiore competenza» (*ibid.*); quest'ultimi, infine, decideranno il revisore a cui affidare la valutazione.

Come si può notare, tale procedura consta di svariati passaggi intermedi, alcuni dei quali sono stati espletati individualmente; è il caso dei Coordinatori, che hanno smistano i prodotti tra i membri del Sub-Gev, e degli Ev, che autonomamente li hanno assegnati ai *referees*. Per comprendere meglio la mole di lavoro che questo tipo di organizzazione comporta, si faccia riferimento alla tabella che illustra la distribuzione dettagliata dei prodotti gestiti. Considerando entrambi i sottogruppi, il numero di contributi per ogni esperto varia da un minimo di 277 prodotti a un massimo di 626, mentre il coordinatore del Sub-Gev di scienze politiche e quello di scienze sociali hanno dovuto gestire, rispettivamente, 2.492 e 3.367[8] prodotti (Tab. 3). In realtà, stando al resoconto del Rapporto d'Area 14, si apprende che «i diversi membri del Gev si sono proposti a gestire certi prodotti e la coordinatrice ha poi confermato le loro scelte e distribuito i prodotti che non erano stati assegnati» (Anvur, 2017a, pp. 23-24). Ciò implica che la Coordinatrice del Gev ha avuto l'onere di controllare e ratificare le richieste relative a tutti i 5.859 prodotti attribuiti all'Area 14.

Ora, volendo accettare l'idea che gli Ev siano in grado di amministrare quantità ingenti di prodotti in ragione di un'ipotetica coerenza di fondo fra le loro competenze e le tematiche generali che caratterizzano i prodotti stessi, non è altresì pensabile che un solo coordinatore possa destreggiarsi in una pletora di alcune migliaia di contributi, padroneggiando al meglio tutte le diverse forme di specializzazione intra-disciplinare in modo da assegnarli sulla base di un principio di coerenza fra i Ssd segnalati nella scheda prodotto e le competenze degli Ev. Tutto ciò potrebbe causare una distribuzione subottimale dei prodotti all'interno del Gev, generando potenziali *bias* il cui controllo risulta alquanto complesso.

[7] Nel caso dell'Area 14, il Gev è stato suddiviso in due Sub-Gev, il primo afferente all'ambito sociologico e il secondo a quello politologico, rispettivamente coordinati da Lorenzo Bordogna e Francesco Tuccari (cfr. Anvur, 2017b, p. 6).

[8] Questi valori derivano da un calcolo additivo operato sul numero di prodotti gestiti riportati nella tabella per ogni Sub-Gev.

 ISBN 9788835103219

Tab. 3 - Organizzazione degli esperti in Sub-Gev, corrispondenti Ssd e distribuzione dei prodotti della ricerca gestiti

Sub-Gev	Ssd	Componenti	Prodotti Gestiti
Scienze Politiche	SPS/01		
	SPS/02	Francesco Tuccari	506
	SPS/03	Attinà Fulvio	385
	SPS/04	Meriggi Marco	382
	SPS/05	Nuti Leopoldo	277
	SPS/06	Galeotti Anna Elisabetta	436
	SPS/13	Piattoni Simona	508
	SPS/14		
Scienze Sociali	SPS/07	Lorenzo Bordogna	581
	SPS/08	Andrini Simona	392
	SPS/09	Baldassarri Delia Stefania	582
	SPS/10	Buzzi Carlo	613
	SPS/11	Colombo Fausto	573
	SPS/12	Wagner Peter	626

Fonte: Anvur, 2017a

Affinché non si verifichino simili condizioni, anzitutto sarebbe utile prestabilire una soglia minima e una massima di prodotti che ogni esperto può ottenere, eventualmente avendo cura di integrare il Gev con nuovi elementi in caso di sovraccarichi dovuti alla scarsa presenza di profili con determinate competenze. In secondo luogo, si ritiene necessario decentralizzare il processo evitando che i soli coordinatori debbano assumersi la responsabilità di vagliare migliaia di prodotti, lasciando che si occupino solo dei casi con candidature multiple, ove le parti coinvolte non riescano ad accordarsi.

3.1.4. Strategie di accountability

La circostanza per cui il medesimo prodotto viene valutato in modo discordante dai due revisori, come si accennava in precedenza, è stato prevalente e ha obbligato il Gev a svolgere un'attività supplementare. Nello specifico, esso ha dovuto avviare una procedura di risoluzione

delle controversie con lo scopo di giungere a una valutazione finale condivisibile e condivisa, costruita sulla base dei punteggi assegnati e dei giudizi espressi dai *referees*. Tale procedura è regolamentata nelle Linee Guida dei Gev: «Nel caso di valutazioni non convergenti […] il sub-Gev crea al suo interno un Gruppo di Consenso con il compito di proporre al Gev il punteggio finale del prodotto oggetto del giudizio difforme dei revisori esterni mediante la metodologia del *consensus report*» (Anvur, 2015b, p. 5). Subito dopo si aggiunge che «Il Gruppo di Consenso può avvalersi anche del giudizio di un terzo esperto» (*ibid*.). Sostanzialmente, queste commissioni di arbitrato[9] avevano facoltà di decidere se dirimere il problema internamente o delegarne la risoluzione ad un revisore esterno, il cui giudizio sarebbe stato definitivo e insindacabile.

Tuttavia, il Gev 14 non sembra aver accolto integralmente le suddette indicazioni. Difatti, nel Rapporto d'Area 14 è scritto che «quando entrambe le valutazioni del prodotto sono state disponibili, il membro Gev ha avuto a disposizione le seguenti azioni: a) convalidare le valutazioni e il voto risultante; b) nel caso in cui abbia ritenuto che ci fosse una discrasia tra giudizio e voto oppure che sussistesse una divergenza marcata tra i due giudizi, ha potuto proporre un cambiamento nella classe finale di merito, c) una terza valutazione, d) un gruppo di consenso» (Anvur, 2017a, p. 25). L'incongruenza con quanto disposto nelle Linee Guida riguarda il fatto di aver considerato la terza revisione come un'opzione alternativa al gruppo di consenso, quando, invece, è chiaramente scritto che si tratta di uno strumento a cui può ricorrere discrezionalmente solo la commissione di arbitrato, e non il singolo membro del Gev.

In realtà, più preoccupante è l'assenza di informazioni circa le modalità di formazione dei gruppi di consenso, di selezione degli esperti che ne hanno fatto parte e di conduzione delle procedure interne atte a sintetizzare i giudizi discordanti.

Permangono dunque alcuni dubbi circa il ricorso ai gruppi di consenso nella Vqr come strumento di risoluzione delle controversie. Anzitutto bisogna considerare un costo inevitabile legato alle diverse afferenze disciplinari dei membri del Gev 14, in cui, com'è noto, ogni

[9] Così vengono anche chiamati i gruppi di consenso nel Rapporto d'Area (Anvur, 2017b, p. 10).

Ssd era rappresentato da un solo Ev (cfr. Tab. 3). In simili circostanze qualsiasi gruppo di consenso derivato dal Gev 14 avrebbe mantenuto la stessa caratteristica e ciò implica, ad esempio, che un prodotto relativo al settore SPS/07 sarebbe stato valutato da una commissione in cui, al massimo, uno solo dei membri avrebbe fatto parte dello stesso settore. Pertanto, sembra lecito chiedersi a che titolo gli altri esperti avrebbero contribuito alla risoluzione della controversia pur non avendo le competenze specifiche necessarie per valutare il prodotto. Diversamente, si potrebbe pensare che il gruppo di consenso non sia stato chiamato a giudicare nuovamente il contributo, bensì a svolgere un lavoro di analisi e interpretazione dei referaggi, cercando di controllare la coerenza dei giudizi scritti e di individuare eventuali fallace argomentative. Tuttavia, il problema sarebbe rimasto nel caso in cui entrambe le valutazioni, per quanto divergenti, avessero rispettato tutti i requisiti necessari.

L'impossibilità di chiarire simili dubbi genera un cono d'ombra in cui si alimenta la sensazione di un'opacità generale che richiama l'esigenza di svolgere una rendicontazione puntuale circa l'applicazione specifica dello strumento. Non basta, ad esempio, scrivere che il gruppo di consenso deve essere «appropriato» (Anvur, 2017b, p. 10) se non si definisce il significato di tale "appropriatezza", così come non è sufficiente dire che gli Ev «hanno potuto proporre un gruppo di consenso per discutere con altri colleghi la valutazione finale» (Anvur, 2017a, p. 71), se si vuole rendere intellegibile e trasparente l'operato di siffatti gruppi.

Alla luce di quanto detto finora, emerge un deficit di *accountability* che tende a presentarsi trasversalmente nelle molteplici fasi che hanno costituito la Vqr. È altresì giusto specificare che l'assenza di una rendicontazione più approfondita delle procedure non implica una loro sostanziale inefficacia, nondimeno impedisce agli stakeholder di comprenderle, valutarle e tesaurizzarle.

Un caso emblematico riguarda l'utilizzo dei giudizi scritti a valle della valutazione di ogni prodotto, che i revisori avevano l'obbligo di formulare per motivare il punteggio assegnato. Tale disposizione rappresentava senz'altro un passo in avanti rispetto alla Vqr 2004-2010, dove invece la compilazione del campo libero riservato ai commenti era facoltativa; ma il fatto di non aver stabilito le coordinate per la costruzione di tale giudizio ha messo a serio rischio la possibilità di analisi comparative utili ai fini di una calibrazione della valutazione.

Per esempio, non è stato specificato il referente del commento, sicché alcuni revisori avrebbero potuto addurre giustificazioni di carattere generale; altri, viceversa, considerare solo due delle tre dimensioni della qualità; altri ancora concentrarsi solamente su una e così via. In definitiva, «La mancanza di vincoli nelle condizioni d'uso non può che influire negativamente sulla completezza e la validità delle informazioni rilevate tramite questo campo, elementi necessari ad assicurare una piena comparabilità dei giudizi e la conseguente utilità ai fini dell'assegnazione del prodotto alla classe di merito finale da parte del Gev» (Fasanella e Di Benedetto, 2014, p. 74)[10].

Concludendo, agli Ev sono state attribuite mansioni fondamentali, fra cui: la redazione dei criteri di valutazione, la nomina dei sub-Gev, la selezione dei revisori e la gestione del rapporto con essi, la creazione dei gruppi di consenso, la predisposizione delle linee guida per i revisori, la scrittura del rapporto finale, la risoluzione delle controversie, l'attribuzione di un prodotto ad un Gev terzo e molte altre (cfr. Anvur, 2015, pp. 17-18). Com'è chiaro, i membri del Gev hanno dovuto far fronte a un'enorme mole di lavoro che implica grandi responsabilità. Spesso, però, anche a causa dei numerosi impegni, sono rimaste zone di discrezionalità troppo ampie che non hanno permesso di operare una rendicontazione adeguata. Per ovviare a questo problema, prima di tutto si rende necessario alleggerire la quantità di lavoro richiesta; e poi, se da un lato non si può pensare di predeterminare minuziosamente il modo in cui i Gev dovrebbero ottemperare ai propri compiti, dall'altro sarebbe opportuno garantire un'*accountability* più solida che renda esplicite le informazioni in grado di spiegare se non il "perché" quantomeno il "come" dei processi attuati.

3.2. I revisori

3.2.1. Il processo di selezione

Ai revisori è affidato il delicato compito di valutare i prodotti e per questo motivo è opportuno curarne il reclutamento con estrema attenzione. Secondo le direttive riportate nelle Linee Guida, i *referees* sono

[10] Per ulteriori approfondimenti sul tema si rimanda al capitolo 4.

stati selezionati «tra gli studiosi e specialisti, tanto italiani quanto stranieri, più autorevoli e scientificamente qualificati delle discipline cui appartengono i prodotti di ricerca da esaminare, scientificamente attivi nel periodo della Vqr. Essi dovranno altresì essere contraddistinti da imparzialità, rigore, equilibrio e senso dei propri limiti» (Anvur, 2015b, p. 4).

In primis, si pone in evidenza la vaghezza della definizione. Quello di "autorevolezza" è un concetto per sua natura sfaccettato, la cui definizione operativa deve avvenire utilizzando indicatori scelti sulla base di fondamenta teoriche esplicite, possibilmente condivise. Un discorso analogo vale per le nozioni di "imparzialità", "rigore", "equilibrio" e "senso dei propri limiti". Diversamente, sarebbe stato più utile individuare una serie di fattori osservabili da introdurre nei criteri di selezione, fra cui, per esempio, l'età, il ruolo accademico e l'affiliazione, che com'è noto sono capaci di incidere sui risultati della valutazione (cfr. Lee *et. al.*, 2013).

A fronte di una definizione così lasca e dell'assenza di dati specifici sulle caratteristiche dei revisori, si può comprendere quanto sia difficoltoso «valutare la rispondenza delle modalità di selezione applicate agli standard prestabiliti, perlomeno non con riferimento al loro esito» (Fasanella e Di Benedetto, 2015, p. 47)[11].

In aggiunta agli ostacoli di tipo teorico-concettuale, ne sono emersi altri di ordine tecnico. Come segnalato nel Rapporto d'Area, «la procedura di invito dei revisori identificata dal CINECA è risultata abbastanza complicata […]. Questo ha portato a un allungamento dei tempi nella definizione dei revisori attivi, e ha disincentivato alcuni potenziali revisori che avevano dato la disponibilità di principio ma che, di fronte alla farraginosità della piattaforma, si sono poi ritirati, soprattutto in certi settori dell'Area 14» (Anvur, 2017a, p. 18). Le defezioni dei potenziali *referees* non sono certo trascurabili, in quanto acuiscono l'incidenza di fattori legati all'autoselezione degli stessi (cfr. Fasanella e Di Benedetto, 2015, p. 47). Infatti, al problema dell'adesione volontaria, si è aggiunto quello della sottorappresentazione dei profili per cui la procedura di adesione, gestita attraverso una piattaforma digitale (CINECA), è risultata particolarmente indigesta. Purtroppo, non vi

[11] Si noti come questa citazione, che nel suo contesto originale si riferisce al primo esercizio valutativo (Vqr 2004-2010), sia ugualmente valida anche per il secondo (Vqr 2011-2014).

sono dati in tal senso, tuttavia, volendo avanzare un'ipotesi, sembrerebbe lecito pensare che, mediamente, i revisori più anziani siano in possesso di competenze informatiche limitate rispetto ai colleghi più giovani, e che questo abbia incrementato la percentuale di rinunce nella corrispettiva fascia d'età. Infine, come si legge, i tassi di abbandono variano molto a seconda dei Ssd, e ciò implica un'ulteriore carenza in termini di rappresentatività per alcune discipline, nonché una scarsa disponibilità di *referees* «in relazione alle competenze richieste» (Anvur, 2017a, p. 70).

Inevitabilmente, tutto ciò ha influenzato in maniera negativa la distribuzione del carico di lavoro, specialmente in settori dove a un alto numero di prodotti non corrispondeva un'adeguata quantità di *referees*. Difatti, «anche se il numero finale dei revisori è risultato sufficiente per concludere la valutazione, […] questi problemi hanno inevitabilmente portato il Gev a utilizzare in modo diseguale i diversi revisori i quali sono stati, in molti casi, chiamati a dover svolgere un numero molto elevato di valutazioni» (Anvur, 2017a, p. 37). Si tratta – è bene rimarcarlo – di un aspetto cruciale poiché l'eventualità di dover svolgere molte revisioni in tempi ristretti contribuisce ad introdurre *bias* aggiuntivi capaci di minare gli esiti della valutazione. A onta di tali considerazioni, nel Rapporto d'Area il problema è stato segnalato solo in via incidentale, senza riportare alcuna informazione più specifica circa la quantità di contributi gestiti dai singoli revisori.

Onde evitare sovraccarichi di lavoro e un'attribuzione dei prodotti mal bilanciata, si potrebbe implementare una procedura che comporti una selezione ragionata dei revisori, ossia legata alla previa conoscenza dei contributi da valutare.

Diversamente, l'attuale gestione del processo di selezione sovverte l'ordine con cui, a giudizio di chi scrive, andrebbero considerati da una parte i prodotti e dall'altra i *referees*. Più precisamente, se è vero che la quantità di *prodotti attesi* può costituire un riferimento utile ai fini della selezione dei revisori, è altresì vero che i *contributi non ancora acquisiti* costituiscono un parametro instabile, soggetto a oscillazioni imprevedibili. Di converso, potrebbe risultare molto più efficace integrare il processo di selezione tenendo in considerazione tutte le informazioni desumibili dai *prodotti effettivi previamente acquisiti*.

Per cui, se nella Vqr 2011-2014 l'individuazione del bacino di *referees* ha preceduto cronologicamente il conferimento dei prodotti, qui

si propone di invertire i due momenti, formando il gruppo dei revisori solo *dopo* aver ottenuto i contributi da sottoporre a valutazione. Tra l'altro, con questa procedura si avrebbe la possibilità di svolgere un'analisi preliminare del materiale pervenuto, tenendo conto sia della numerosità sia, specialmente, delle caratterizzazioni tematiche più o meno rappresentate. Così, in una fase successiva, considerando i dati emersi in sede di analisi, si potrebbe stabilire da un lato la quantità ottimale di revisori da impiegare, dall'altro il dominio di competenze che essi dovrebbero coprire sulla base dell'eterogeneità disciplinare che contraddistingue i prodotti.

D'altronde, già nell'ultima Vqr è emersa l'esigenza di integrare la lista dei revisori con nuove acquisizioni effettuate *in itinere*, vale a dire in un momento *successivo* al conferimento dei prodotti. Nelle Linee Guida, ad esempio, è indicato che «Il processo di integrazione della lista [dei revisori] continuerà per tutta la durata della valutazione, sulla base delle necessità che dovessero emergere a valle della trasmissione dei prodotti da parte delle Istituzioni» (Anvur, 2015b, p. 5); e ancora: «Nel corso della fase di valutazione che è iniziata a fine maggio, il numero dei revisori a disposizione del Gev, è progressivamente aumentato» (Anvur, 2017a, p. 18).

Ora, siccome l'idea di selezionare i *referees* in base a necessità accertate non appare del tutto peregrina, perché non applicarla a monte del processo di selezione? Un sistema così implementato aiuterebbe ad ottenere un numero sufficiente di revisori sia in senso assoluto, sia in proporzione alle diverse aree disciplinari cui i prodotti afferiscono. Inoltre, evitando eccessivi aggiustamenti *in itinere* della lista dei revisori si contengono i rischi «di un abbassamento degli standard in relazione alle necessità pratiche e alle tempistiche dell'esercizio di valutazione» (Fasanella e Di Benedetto, 2015, p. 48). A dire il vero, i suggerimenti del Gev 14 sembrano andare in direzione opposta, allorché si auspica «di non disperdere l'esperienza accumulata e la lista di revisori ottenuta, ma di sottoporla a verifica con aggiornamenti costanti così che si arrivi alla nuova Vqr con una lista *già pronta* di revisori» (Anvur, 2017a, p. 71, corsivo aggiunto).

 ISBN 9788835103219

3.2.2. Il matching *prodotto-revisore*

La soluzione prospettata alla fine del precedente paragrafo offre molteplici vantaggi anche in relazione al fondamentale problema del *matching* prodotto-revisore[12]; ma prima di chiarire come possa essere proficuamente applicata è bene avanzare alcune brevi considerazioni sulla procedura di assegnazione.

Ogni *referee* esercita il proprio lavoro attingendo al bagaglio di competenze sviluppato nel corso della sua carriera. In seno alla Vqr, questo patrimonio di saperi può e deve essere capitalizzato, tuttavia, per riuscirci bisogna garantirne la massima compatibilità in relazione al contenuto dei prodotti da valutare. Pertanto, si rende necessario seguire uno schema di attribuzione in grado di favorire la più alta coerenza fra gli argomenti trattati nei prodotti e le specializzazioni disciplinari dei revisori. Ciò, inoltre, concorre a migliorare la qualità delle valutazioni tenendo sotto controllo l'intervento di elementi distorsivi indesiderati.

Per questi motivi, la distribuzione dei prodotti ai *referees* – che spetta al Gev (cfr. § 3.1.3) – è stata operata prendendo come riferimento gli Erc e i Ssd, in modo da rispettare il principio di coerenza appena descritto. Purtroppo, però, almeno nell'ambito dell'Area 14, queste categorizzazioni si sono dimostrate troppo vaste ed eterogenee, con la conseguenza che gli Ev si sono trovati in seria difficoltà, dovendo «distribuire per la valutazione prodotti di cui non sono personalmente esperti e di cui non conoscono potenziali valutatori competenti sul tema» (Anvur, 2017a, p. 70). In alternativa si è tentato di ricorrere allo strumento delle parole-chiave, segnalate dagli autori al fine di indicare le tematiche salienti del prodotto. Nondimeno, il loro utilizzo si è rivelato tutt'altro che risolutivo. Esse, infatti, «sono scelte dagli addetti stessi che, a volte, indicano solo il settore disciplinare o sua sezione, e quindi non sono per niente informativi, a volte forniscono liste lunghissime non sempre rispondenti alla loro stretta competenza di ricerca» (*ibid*.).

Si è così delineata una situazione in cui alcuni revisori sono stati costretti a leggere e giudicare i prodotti senza avere la concreta possibilità di comprenderne appieno l'originalità, il rigore procedurale e l'impatto potenziale. Nella migliore delle ipotesi, in simili condizioni

[12] Con questa locuzione ci si riferisce al procedimento che regola l'assegnazione dei contributi ai singoli revisori.

il *referee* ha rifiutato la revisione, concorrendo allo slittamento dei tempi di chiusura dell'esercizio dovuto all'esigenza di allocare nuovamente il prodotto. Volendo essere più pessimisti, non si può escludere che qualcuno abbia proceduto comunque alla valutazione dei contributi pur non essendo in possesso delle opportune competenze disciplinari. Chiaramente, un simile atteggiamento va disincentivato con forza, in virtù delle sue ripercussioni ferali sugli esiti della valutazione.

Osservando la tabella che riassume le diverse motivazioni che hanno portato i revisori a rifiutare la valutazione dei prodotti, si nota che su 1.511 rifiuti ben 606 si devono alla mancanza di competenze, vale a dire più di un terzo del totale (Tab. 4). Per inciso, la seconda giustificazione in termini di frequenza riguarda la mancanza del tempo necessario per valutare, il che fa pensare all'esigenza di una più oculata gestione delle tempistiche. Ad ogni modo, prendendo in considerazione questi dati appare chiaro che il *matching* prodotto-revisore, così come implementato nell'ultima Vqr, ha prodotto un'assegnazione subottimale dei contributi. È dunque indispensabile promuovere l'introduzione di misure correttive che tendano a migliorare l'efficacia del *matching*, massimizzando la sovrapponibilità del contenuto dei prodotti con il *background* teorico-pratico dei *referees*.

Tab. 4 - Motivo del rifiuto dei revisori italiani e stranieri (val. ass. e %)

Ho già abbastanza da valutare	Non comprendo la lingua	Non dispongo del tempo necessario per valutare	Non posseggo le competenze necessarie per valutare	Sono in conflitto di interessi	Altro	Totale
271	79	302	606	118	135	1.511
18	5,2	20	40,1	7,8	8,9	100

Fonte: Anvur, 2017a

Poc'anzi si accennava all'utilità di presentare una soluzione analoga a quella proposta per l'assegnazione dei contributi agli Ev. Ebbene, nello specifico si tratterebbe di adottare un sistema di classificazione più raffinato rispetto ai Ssd e agli Erc, basato sulla formulazione di un vasto repertorio di etichette disciplinari da assegnare a tutti i prodotti (cfr. Fasanella e Di Benedetto, 2015, p. 57). Successivamente, le stesse etichette andrebbero attribuite ai revisori o tramite un'auto-di-

chiarazione degli stessi o, in maniera centralizzata e meglio controllabile, mediante l'analisi delle loro più recenti pubblicazioni (ad esempio le ultime dieci, oppure quelle relative ai tre anni di attività che precedono la Vqr). Così facendo, la strategia di *matching* assumerebbe un più alto livello di precisione, in quanto capace di operare su gruppi di prodotti e di revisori che condividono la medesima etichetta.

Prima di procedere oltre, è utile soffermarsi su qualche annotazione fondamentale. Anzitutto, assumendo un'istanza di partecipazione democratica alla valutazione (cfr. Palumbo e Torrigiani, 2009) – ma anche di funzionalità dello strumento – è auspicabile che il direttivo dell'Anvur, attraverso un percorso di negoziazione, condivida la formulazione del *thesaurus* di etichette con tutte le comunità scientifiche coinvolte. Inoltre, si evidenzia che la soluzione suggerita ha il vantaggio non trascurabile di sollevare i membri del Gev dall'onere di gestire l'intera opera di smistamento dei prodotti. Nondimeno, è d'obbligo richiamare la perdurante centralità del ruolo degli Ev, ai quali comunque spetterebbe la selezione finale del singolo revisore a cui assegnare il contributo, scelto nel bacino dei candidati idonei suggeriti da un algoritmo che gestisce la prima fase del *matching*. Sarebbe bene, oltretutto, che gli Ev mantenessero la possibilità di attribuire i prodotti anche a *referees* esterni al gruppo, purché sussistano motivazioni ragionevoli e che queste vengano esplicitate. Infine, in sede di programmazione della piattaforma atta a computare le varie combinazioni prodotto-revisore, è indispensabile includere i dati necessari per tenere conto delle situazioni conflittuali, specialmente se i criteri sono difficili da controllare, come nel caso di «prodotti di cui siano autori o co-autori coniugi, parenti o affini fino al 4° grado» o «presentati da università presso cui i membri stessi, o i revisori da loro scelti, abbiano o abbiano avuto un rapporto di lavoro o con le quali abbiano svolto incarichi o collaborazioni ufficiali» (Anvur, 2017b, p. 12).

*3.2.3. L'***accountability** ***recitativa***

In considerazione della sua centralità, anche il lavoro svolto dai revisori deve essere adeguatamente rendicontato. Tuttavia, per poterlo fare bisogna anzitutto avere ben chiaro che cosa si debba intendere con il termine "*accountability*". In secondo luogo, è altresì necessario

strutturare le procedure in modo che la loro realizzazione risulti chiara e definita, tanto agli occhi degli attori direttamente coinvolti quanto a quelli degli osservatori terzi.

Non si può sottacere, allora, il fatto che nella Vqr 2011-2014 queste due istanze abbiano trovato un riscontro solo parziale. Facendo riferimento alla seconda, può essere d'aiuto prendere come esempio il caso dell'*informed peer review*. Con questa locuzione si vuole indicare un metodo in base al quale «ai revisori sono fornite, oltre al testo pdf del prodotto, le informazioni contenute nella scheda prodotto compilata dall'addetto cui il prodotto è associato. Le informazioni contenute nella scheda prodotto riguardano, tra le altre cose, gli indici bibliometrici, laddove presenti, la classificazione della rivista, nel caso di articolo in rivista, e il superamento di una valutazione *peer review* in caso di monografia o saggio in volume» (Anvur, 2017b, p. 9).

Se, da un lato, si tratta di uno strumento potenzialmente adatto a sostenere i *referees* nella formulazione dei punteggi e dei giudizi da attribuire ai prodotti, dall'altro non vi sono indicazioni sufficienti circa le modalità con cui essi avrebbero dovuto utilizzare informazioni contenute nella scheda prodotto. Che peso assegnare all'eventualità che un articolo abbia già superato o meno una valutazione *peer*? In che modo la classificazione della rivista nobilita ovvero scredita il valore del prodotto che contiene? Il revisore scrupoloso che si fosse posto queste domande avrebbe dovuto fare l'ulteriore sforzo di arrendersi all'assenza di risposte.

In aggiunta, in appendice al Rapporto si può leggere che i dati integrativi dell'*informed peer review* «non predeterminano la valutazione che è responsabilità del revisore» (*ibid.*). Tuttavia, la risoluzione dei casi discordanti passa esplicitamente attraverso la valutazione della misura in cui il revisore abbia «tenuto conto delle informazioni della scheda prodotto» (Anvur, 2017a, p. 23) o vi sia «la mancata considerazione dei dati concernenti la sede di pubblicazione» (*ivi*, p. 71). Pertanto, da un lato, sembrerebbe che le informazioni aggiuntive non *pre*-determinino la valutazione, dall'altro, però, in qualche modo la devono determinare; a questo punto è chiaro che il fatto che lo facciano *ex post*, *in itinere* o *ex ante* non cambia poi molto. In aggiunta, alcune indicazioni risultano sommarie: che cosa significa, per esempio, "tenere conto" delle informazioni? Sicuramente che i dati vanno osservati poiché giocano un ruolo nella valutazione globale, ma in che modo? Poniamo il caso che

un revisore tenga in massima considerazione una certa rivista cui appartiene il prodotto che sta giudicando; egli potrebbe "tenere conto" di questo in modo da sovrastimare l'importanza di tale informazione rispetto alle altre, giungendo così a una valutazione distorta.

Alla luce di quanto detto, si può comprendere meglio l'importanza di fornire indicazioni più precise al fine di armonizzare il comportamento dei revisori in sede di valutazione, e, in generale, di redigere un *corpus* di linee guida adeguato, facendo sì che le direttive possano essere attualizzate nel modo più agile e trasparente possibile. Infine, si rimarca il fatto che le linee guida non hanno come obiettivo quello di contrastare *tutti* gli elementi di soggettività connaturati alla Vqr; alcuni di essi rappresentano una risorsa da valorizzare, altri, invece, un fattore distorsivo indesiderato. La loro efficacia, allora, dipende dalla capacità di discernere tali componenti, eradicando quelle che si ripercuotono negativamente sulla validità degli esiti del processo valutativo.

Ora, per quanto concerne la seconda istanza, ossia la specificazione del concetto di *accountability*, è il caso di riportare la definizione di Martini e Cais, ritenuta da chi scrive particolarmente funzionale per il prosieguo del discorso. Secondo gli autori, l'*accountability* si riferisce al «dovere che un soggetto responsabile di un'organizzazione (o di una politica, di un progetto) ha di "render conto" a particolari interlocutori esterni delle scelte fatte, delle attività e dei risultati di cui è autore o responsabile» (Martini e Cais, 2000, p. 410).

Lasciando un attimo da parte ulteriori considerazioni, al momento si ritiene opportuno spostare l'attenzione su ciò che, in via ipotetica, avrebbe dovuto costituire un'operazione di *accountability*.

Come era stato preannunciato (cfr. Anvur, 2017b, p. 9), l'Anvur ha provveduto alla pubblicazione dell'elenco dei revisori, ossia una lunghissima lista di 11.750[13] nominativi priva di qualsiasi altra informazione. Non è presente né l'indicazione della sede di affiliazione (se estera o italiana), né il settore scientifico-disciplinare, il che impedisce di risalire al numero preciso di revisori che hanno operato per ogni macro-area, tantomeno per ogni Ssd. Invero, armandosi di buona volontà, è possibile risalire a questi dati attraverso una ricerca per nome e cognome sfruttando l'anagrafica dei docenti universitari gestita dal

[13] Si evidenzia, tra l'altro, una discrepanza fra il numero dei revisori inclusi nella lista e quello riportato nel rapporto generale, pari a 12.731 (cfr. Anvur, 2017, p. 26).

CINECA, che include numerose variabili individuali come la qualifica, il genere, il settore scientifico-disciplinare, il settore concorsuale, l'ateneo, il dipartimento di affiliazione, ecc. Purtuttavia, non è possibile effettuare la stessa operazione per i revisori che hanno sede all'estero[14], i quali, dunque, non rientrano nell'analisi.

Purtroppo, anche focalizzando l'attenzione solo sui casi con affiliazione italiana, sorge un altro problema legato al fatto che all'interno dell'organico nazionale dei docenti universitari sono presenti alcune omonimie che, in assenza di dati aggiuntivi oltre al nome e al cognome, non possono essere disambiguate. In sostanza, può accadere che cercando un revisore all'interno dell'anagrafica si ottengano più persone, rendendo impossibile riuscire a identificare quali docenti siano stati effettivamente impegnati nella Vqr.

Così, a seguito delle inevitabili scremature, si ottengono 8.149 nominativi per cui è possibile ricavare informazioni utili ai fini dell'analisi. In particolare, dal settore scientifico-disciplinare si desumono il settore concorsuale e l'area Cun di afferenza, che permettono di ottenere la Tabella 5[15]. Per ogni area disciplinare sono indicati il numero di revisori che hanno preso parte alla Vqr e quello di tutti i docenti a livello nazionale (entrambi in valori assoluti e percentuali). Ciò ha permesso di calcolare la quota di revisori sul totale dei docenti (Tab. 5).

Con esclusivo riferimento all'Area 14, risulta alquanto difficile ricavare dati supplementari che consentano elaborazioni più significative. Nel Rapporto d'Area è riportato soltanto il numero di revisori (distribuiti per diverse altre informazioni, come la sede di afferenza e il Ssd) ripetuti per ogni Ssd di competenza e per il numero di revisioni effettuate[16]. Insomma, si tratta di dati pressoché inservibili. Il massimo che si è potuto fare è una stima del carico di lavoro medio dei revisori

[14] A meno di una ulteriore ed onerosa analisi svolta su database esteri.

[15] Per inciso, si noti che la tabella riporta anche la distribuzione dell'organico nazionale per area disciplinare. Ebbene, in una situazione ideale, le due distribuzioni dovrebbero presentare percentuali simili, tuttavia è possibile notare come in alcuni casi vi sia un certo squilibrio. La sovra-rappresentazione di alcune aree rispetto alla situazione nazionale stupisce, poiché si tratta di aree disciplinari in cui non viene normalmente applicata la *peer review* e per le quali la valutazione tra pari rappresentava una sperimentazione; in altre parole, ci si sarebbe aspettati una situazione in cui le aree non-bibliometriche fossero numericamente maggiori rispetto a quelle bibliometriche.

[16] A complicare le cose, va fatto notare che il numero di revisori ripetuti per ogni Ssd di competenza cambia tra il Rapporto di Area 14 e il Rapporto Finale Anvur.

per settore scientifico-disciplinare. Nello specifico, la tabella riporta i valori assoluti e percentuali dei revisori afferenti a ogni Ssd dell'Area 14 impegnati nella Vqr, nonché la quantità di prodotti inviati a valutazione relativi al medesimo settore. In tal modo è stato possibile dare una parziale idea del carico di lavoro svolto dai *referees* calcolando il rapporto fra prodotti conferiti e revisori disponibili (Tab. 6).

Tab. 5 - Distribuzione dei revisori per area disciplinare, distribuzione dell'organico nazionale per area disciplinare e percentuale dei revisori sul totale nazionale di area

Area disciplinare	Frequenze assolute revisori	% revisori	Frequenze assolute nazionale	% nazionale	% revisori su totale nazionale
01 - Scienze matematiche e informatiche	144	1,8	3.026	5,5	4,8
02 - Scienze fisiche	326	4	2.162	4	15,1
03 - Scienze chimiche	486	6	2.796	5,1	17,4
04 - Scienze della terra	200	2,5	1.002	1,8	20,0
05 - Scienze biologiche	480	5,9	4.626	8,5	10,4
06 - Scienze mediche	942	11,6	9.185	16,8	10,3
07 - Scienze agrarie e veterinarie	496	6,1	2.964	5,4	16,7
08 - Ingegneria civile e architettura	414	5,1	3.415	6,3	12,1
09 - Ingegneria industriale e dell'informazione	232	2,8	5.303	9,7	4,4
10 - Scienze dell'antichità, filologico-letterarie e storico-artistiche	1.292	15,9	4.725	8,7	27,3
11 - Scienze storiche, filosofiche, pedagogiche e psicologiche	1.075	13,2	4.301	7,9	25,0
12 - Scienze giuridiche	1.314	16,1	4.635	8,5	28,3
13 - Scienze economiche e statistiche	343	4,2	4.775	8,7	7,2
14 - Scienze politiche e sociali	405	5	1.669	3,1	24,3
Totale	8.149	100	54.584	100	14,9

In fin dei conti, il vero rammarico deriva dalla consapevolezza che per gli addetti ai lavori la pubblicazione di simili tabelle risulterebbe spesso agevole; per esempio, disponendo di informazioni utili che riguardano i revisori, come l'ateneo di affiliazione, la posizione accademica, il genere, il Ssd, il numero di prodotti referati ecc., perché non

organizzarle in una matrice pubblica, accessibile a tutti, in modo da rafforzare la trasparenza e al contempo offrire un prezioso strumento d'analisi? Si noti bene che un'iniziativa simile permetterebbe comunque di conservare l'anonimato dei revisori, omettendo il dato sui singoli prodotti valutati da ognuno.

Tab. 6 - Distribuzione dei revisori di area 14, distribuzione dei prodotti conferiti e rapporto tra prodotti conferiti e revisori per Ssd

	Revisori		Nazionale				
Ssd	Frequenze assolute	Frequenze percentuali	Frequenze assolute	Frequenze percentuali	% revisori su totale Ssd nazionale	Prodotti conferiti	Rapporto prodotti conferiti/ revisori
SPS/01	25	6,2	111	6,7	22,5	187	7,5
SPS/02	32	7,9	118	7,1	27,1	217	6,8
SPS/03	22	5,4	62	3,7	35,5	114	5,2
SPS/04	78	19,3	219	13,1	35,6	410	5,3
SPS/05	8	2,0	22	1,3	36,4	40	5
SPS/06	19	4,7	72	4,3	26,4	120	6,3
SPS/07	72	17,8	386	23,1	18,7	695	9,7
SPS/08	61	15,1	309	18,5	19,7	516	8,5
SPS/09	33	8,1	139	8,3	23,7	269	8,2
SPS/10	12	3	71	4,3	16,9	131	10,9
SPS/11	12	3	49	2,9	24,5	75	6,3
SPS/12	15	3,7	60	3,6	25	103	6,9
SPS/13	5	1,2	27	1,6	18,5	51	10,2
SPS/14	11	2,7	24	1,4	45,8	43	3,9
Totale	405	100	1.669	100	24,3	2971	7,3

Al là del suo limitato potere esplicativo, questa breve analisi ha avuto lo scopo di mettere in luce la profonda differenza tra la concezione di *accountability* propugnata in questa sede e quella che talvolta traspare dai documenti pubblici incaricati di rendicontare il processo valutativo. Ne è un esempio il primitivo elenco dei nomi dei *referees*,

incapace di dire alcunché circa le "scelte fatte", le "attività" e i "risultati"[17] di cui gli ideatori della Vqr sono responsabili. E intanto – chi è impegnato in questo campo d'indagine lo sa bene – è sempre più difficile portare avanti studi e ricerche, poiché le informazioni attinenti alle procedure svolte sono sì, in costante aumento, ma spesso anche qualitativamente esili. Di converso, è sempre più probabile che i ricercatori siano costretti a recuperare faticosamente e a mettere insieme pezzi di informazione che in realtà sono già sistematizzati presso qualche *database* centrale. Tutto ciò contribuisce a creare lo spettro di un'*accountability* che ostacola la vera rendicontazione; un'*accountability* che si presume sostanziale e al contempo mal cela la sua anima recitativa.

3.3. Conclusioni

Malgrado il presente lavoro abbia posto l'accento sulle principali problematiche che hanno afflitto la Vqr, e in particolare l'operato dei Gev e dei revisori, certamente vi sono stati anche degli elementi positivi o, quantomeno, migliorativi rispetto al precedente esercizio valutativo. A titolo d'esempio, se nella prima Vqr «almeno con riferimento ai criteri di valutazione generale, non vi è stata alcuna consultazione della comunità scientifica» (Di Benedetto, 2015, p. 100), nella seconda «Entrambi i bandi sono però stati pubblicati in bozza prima della definizione dei criteri, al fine di permettere uno scambio con la comunità scientifica» (*ibid*., testo in nota). Nondimeno, rimane il fatto che i criteri per la definizione del concetto di qualità «sono stati determinati dal decreto ministeriale» (*ibid*., testo in nota) senza la partecipazione di attori terzi.

Tale situazione permette di introdurre un tema cruciale in quanto costituisce l'emblema delle attuali forme di coinvolgimento della comunità scientifica nell'esercizio della Vqr. Fortunatamente, da un certo punto di vista si sta affermando la consapevolezza circa l'importanza del contributo che gli stakeholder possono offrire. In questo senso, alcuni segnali emergono nell'introduzione del Rapporto d'Area, dove si legge che «Un simile risultato non sarebbe stato in alcun

[17] Si veda nuovamente la definizione di Martini e Cais, a cui si fa riferimento.

 ISBN 9788835103219

modo raggiungibile senza la collaborazione di tutta la comunità scientifica principalmente dell'area 14, ma non solo, e dei *referees* italiani e stranieri che hanno dato la loro disponibilità a valutare i prodotti di ricerca e che hanno effettuato le valutazioni negli stretti limiti di tempo concessi dalla procedura di valutazione» (Anvur, 2017a, p. 14).

A ben vedere, però, si tratta di un riconoscimento che chiama in causa uno sforzo straordinario e contingente, uno sforzo non già ricompreso in un disegno sistemico, bensì circoscritto al solo svolgimento pratico della valutazione. Diversamente si ritiene che questa collaborazione, ancorché apprezzata, non possa essere parziale; non possa, cioè, esercitarsi esclusivamente *in itinere*, ma è necessario che sia prevista in tutte le fasi della Vqr, dalla stipulazione del concetto di qualità alla restituzione degli esiti.

Pertanto, è fondamentale programmare i prossimi esercizi valutativi facendo tesoro dell'esperienza maturata dalla comunità scientifica – che della Vqr è protagonista – includendola a vario titolo nella progettazione delle procedure che la coinvolgono. La speranza è che tali suggerimenti possano essere accolti, e in effetti qualche proposta interessante già è stata avanzata in vista del prossimo esercizio valutativo. Ad esempio, l'Anvur suggerisce di costituire sin da subito dei gruppi di studio di area che includano tra le varie figure anche alcuni membri dei precedenti Gev, con l'obiettivo di analizzare il lavoro svolto durante la Vqr e proporre interventi correttivi sulla base dei problemi emersi. Ciò permetterebbe di «passare il testimone da un Gev all'altro facendo tesoro del patrimonio di conoscenze "sul campo" accumulato nel presente processo di valutazione» e di «fornire una lista dei problemi concreti che si sono incontrati in corso d'opera, e delle soluzioni elaborate dopo un processo di *trials and errors*, nonché degli auspicabili miglioramenti a queste soluzioni» (Anvur, 2017a, p. 72).

In ogni caso, va ritenuto che lo scopo ultimo dovrebbe consistere nell'imperniare la Vqr attorno a un percorso condiviso di partecipazione, capace di coinvolgere tutti gli stakeholder e avviare pratiche di co-costruzione dei programmi nella piena trasparenza dei processi decisionali. Solo così è possibile pervenire a una valorizzazione delle variabili umane in gioco, che questa attività, sperabilmente, non potrà mai trascendere del tutto.

 ISBN 9788835103219

4. La scheda di valutazione dei prodotti scientifici

di Federica Floridi

4.1. Introduzione

L'esigenza di trasparenza e di *accountability* della Vqr 2011-2014 si pone anche per gli strumenti e le tecniche utilizzati per rilevare empiricamente la qualità dei prodotti scientifici. La ricostruzione di cosa sia avvenuto a tale livello non risponde solamente ad una esigenza metodologica ma rappresenta un passaggio fondamentale per comprendere gli esiti dell'esercizio valutativo.

Il ruolo che ricopre la scheda di valutazione all'interno del più ampio processo della Vqr è di estrema delicatezza e rilievo poiché è in questa fase che – grazie all'applicazione di tale strumento – si concretizza il passaggio della nozione di qualità della ricerca da concetto astratto e generale a dato empirico e osservabile (Fasanella e Martire, 2017, p. 91).

Idealmente, possiamo immaginare che la costruzione di uno strumento di rilevazione possa essere ricompresa nel più ampio processo che porta dai concetti astratti agli indici empirici e che si articola – seguendo il noto paradigma di Lazarsfeld – in quattro fasi: rappresentazione figurata del concetto; specificazione delle dimensioni; scelta degli indicatori osservabili; ricomposizione degli indicatori in un indice sintetico. «Ci si può evidentemente chiedere a che cosa serva prendere la via indiretta delle dimensioni, degli indicatori e delle misure, dato che alla fine si arriva a semplici descrizioni verbali. La risposta è semplice. L'analisi nel suo insieme riesce a semplificare la rappresentazione del concetto primitivo, in modo da ottenere un accordo intersoggettivo sul suo contenuto» (Boudon e Lazarsfeld, 1965; trad. it. a c. di Cavazzani, 1969, p. 19).

 ISBN 9788835103219

Nel caso della scheda di valutazione della Vqr 2011-2014 dalla definizione del concetto di qualità della ricerca – declinato attraverso tre dimensioni costitutive – alla costruzione della scheda di valutazione non vi è stato alcun passaggio intermedio di individuazione di indicatori prossimi al livello empirico. In pratica, è stato compiuto un salto logico-procedurale dalle dimensioni del concetto alla formulazione delle domande, il cui testo corrisponde alla definizione stipulativa delle suddette dimensioni. L'assenza di operazioni di chiarificazione concettuale e di specificazione di significato del concetto di qualità della ricerca potrebbero aver generato delle difficoltà e in alcuni casi aver prodotto distorsioni nella rilevazione del concetto di qualità della ricerca.

4.2. La struttura della scheda Vqr 2011-2014

Uno strumento di rilevazione può risultare perfettamente adeguato dal punto di vista del contenuto semantico ma scarsamente utilizzabile o, viceversa, di agevole applicazione ma scarsamente rilevante dal punto di vista del contenuto semantico. La scheda, dunque, doveva conciliare due necessità ugualmente avvertite. Da una parte, doveva essere in grado di riprodurre la definizione di qualità della ricerca così come stabilita dal d.m. n. 458 del 2015 e adottata dall'Anvur per l'esercizio valutativo 2011-2014, in modo tale da produrre proprio il significato che si intendeva rilevare. Dall'altra, doveva risultare sufficientemente governabile da parte dei *referees*, così da consentire l'espressione di un giudizio valutativo in merito alla qualità scientifica dei prodotti di ricerca.

Prima di proseguire nella descrizione, occorre premettere che la procedura messa in atto dalla Vqr per la compilazione della scheda di valutazione dei prodotti è avvenuta attraverso il supporto informatico e gli stessi prodotti scientifici sono stati forniti in formato pdf. I revisori sono stati contattati tramite posta elettronica e, in assenza di conflitti d'interesse, potevano accedere a una piattaforma digitale nella quale avveniva la vera e propria valutazione attraverso la compilazione dell'apposita scheda.

Partendo, dunque, dalla definizione del bando e dalle linee guida della Vqr, da cui sono stati individuati i tre criteri costitutivi della qualità della ricerca, la scheda è stata strutturata in cinque sezioni.

 ISBN 9788835103219

Scheda di valutazione dei prodotti 2011-2014 Area 14

Criterio	Valutazione
Q1. Originalità: ha lo scopo di misurare quanto sia innovativo il prodotto di ricerca, rispetto a un nuovo modo di pensare, nuove prospettive, tesi, temi e/o fonti, in relazione all'oggetto scientifico della ricerca, e pertanto quanto si distingua dai precedenti lavori sullo stesso tema. Assegni un punteggio da 1 (valore minimo) a 10 (valore massimo) all'originalità del prodotto.	
Q2. Rigore metodologico: ha lo scopo di misurare: i) il livello di chiarezza con cui il prodotto presenta gli obiettivi di ricerca; ii) il livello di competenza scientifica e di padronanza dello stato dell'arte; iii) la capacità di adottare una metodologia appropriata rispetto all'oggetto della ricerca; iv) il raggiungimento (realizzazione) degli obiettivi prefissi. Assegni un punteggio da 1 (valore minimo) a 10 (valore massimo) al rigore metodologico del prodotto.	
Q3. Impatto attestato o potenziale: ha lo scopo di misurare il livello rispetto a cui il prodotto ha esercitato – o è presumibile eserciti in futuro – un'influenza sulla comunità scientifica presente, anche in base alla sua capacità di rispettare standard internazionali di qualità della ricerca.	
Totale e classe di merito	
Q4. Formulazione (campo libero) di un giudizio sintetico finale (obbligatorio, con un minimo di 50 parole e un massimo di 200):	

Fonte: Anvur, 2017d

Le prime tre sono state predisposte per poter esprimere un giudizio quantitativo tramite l'assegnazione di un punteggio in una scala da 1 (valore minimo) a 10 (valore massimo) su ognuno dei tre criteri: originalità, rigore metodologico e impatto attestato o potenziale. La quarta sezione, invece, permetteva di visualizzare la classe di merito finale nella quale ricadeva il prodotto – eccellente, elevato, discreto, accettabile, limitato – in base alla somma dei punteggi ottenuti sui tre criteri. Dopo aver visualizzato la classe di merito, la scheda consentiva ai revisori di modificare i punteggi precedentemente assegnati ai tre criteri al fine di un'eventuale ricollocazione del prodotto in una classe di merito differente. Infine, l'ultima sezione era dedicata ad un campo libero per esprimere un giudizio sintetico finale obbligatorio, che potesse in qualche modo argomentare l'assegnazione del prodotto alla classe di merito (Anvur, 2017c, p. 5).

Il testo delle prime tre domande, Q1-Q3, mostra un'ambiguità di fondo poiché fa esplicito riferimento alla pretesa di poter misurare i

tre criteri di qualità della ricerca, nonostante sia necessario rispettare alcuni requisiti affinché si possa debitamente parlare di misurazione «[…] molti ricercatori usano liberamente i termini "misurare/misura/misurazione" per ogni procedimento di registrazione di stati (ordinamento, conteggio, *scaling*, e magari anche per la classificazione). Questo vero e proprio abuso terminologico non ha altro motivo che l'ansia di legittimazione scientifica» (Marradi, 2007, p. 144).

In effetti, l'attribuzione di un punteggio alle dimensioni della qualità della ricerca si configura come un'operazione di *scaling* piuttosto che di misurazione, la quale dovrebbe: a) rilevare una proprietà continua, b) basarsi su un'unità di misura e c) poter confrontare l'unità di misura e lo stato dei vari oggetti (o soggetti) sulla proprietà che si sta misurando[1].

Ogni criterio è stato, quindi, presentato nella scheda sotto forma di uno stimolo al quale assegnare un punteggio seguendo uno *scaling* tipo Cantril. La caratteristica peculiare della scala Cantril, che venne ideata dall'omonimo autore (Cantril, 1965), risiede nell'essere una *self anchoring scale*, ossia una scala ancorata semanticamente ai poli estremi sulla base delle percezioni, dei valori e delle assunzioni dell'intervistato stesso, al quale viene chiesto di posizionarsi rispetto al presente, al passato e al futuro (Di Franco, 1989, p. 60) lungo un *continuum* che va da 1 – giudizio fortemente negativo – a 10 – giudizio fortemente positivo –, restituendo in tal modo variabili quasi-cardinali (Pitrone e Pavisc, 2003, p. 130) che, generalmente, vengono trattate come numeri naturali.

L'utilizzo di una scala Cantril come scala di valutazione ha sicuramente dei vantaggi rilevanti, tra i quali il principale è quello di far riferimento al sistema di valutazione scolastico italiano. Pertanto, possiamo considerare ottimale la sua applicazione a una popolazione scolarizzata in Italia (Poli, 2008, p. 55; Pitrone e Pavisc, 2003, p. 135). Tale caratteristica permette di socializzare agevolmente lo strumento, il quale, almeno per i revisori italiani, dovrebbe risultare maggiormente rispondente ai sistemi di valutazione conosciuti e utilizzati (Fasanella e Di Benedetto, 2014 e 2015). Inoltre, ottenendo dei punteggi da 1 a 10 è facilmente ricomponibile/calcolabile un indice sintetico costruito per via matematica ed esprimibile tramite un unico punteggio numerico.

[1] Per una trattazione esaustiva si rimanda a Marradi (2010).

Senza dubbio, sul piano della sensibilità dello strumento di rilevazione, sono stati apportati dei significativi cambiamenti rispetto alla precedente edizione della Vqr che presentava solamente quattro alternative di risposta, peraltro con una formulazione molto problematica (Fasanella e Di Benedetto, 2015, p. 58). Tuttavia, occorre riflettere anche su alcune distorsioni sistematiche che possono essere generate dall'utilizzo di un tale strumento di rilevazione.

Innanzitutto, è stata spesso rilevata una tendenza a posizionarsi sui punteggi centrali, e ciò in funzione dell'eccezionalità che rappresentano i punteggi estremi della Cantril (Pitrone e Pavisc, 2003, p. 135). Che tale fenomeno si sia verificato nella Vqr 2011-2014 è confermato, ad esempio, dall'estratto del rapporto Vqr in cui si afferma che «Le valutazioni di merito sono in generale (…) particolarmente parsimoniose nella sub-area sociologica dove i prodotti eccellenti non superano il 4,7%» (Anvur, 2017a, p. 41). Tale fenomeno distorsivo riduce, di fatto, la sensibilità della scala adottata per formulare la valutazione.

Inoltre vi è il rischio di incorrere nell'assegnazione discrezionale dei punteggi in quei casi in cui si inneschino effetti di reazione all'oggetto (Poli, 2008; Pitrone e Pavisc, 2003) che induce ad assegnare i punteggi facendo riferimento alle caratteristiche note del prodotto – ad esempio l'autore, l'affiliazione, l'approccio utilizzato, ecc. – piuttosto che al prodotto stesso. Considerando che i revisori conoscevano l'identità degli autori e le caratteristiche dei prodotti sottoposti a valutazione, il verificarsi di tale effetto distorsivo è molto plausibile: una parte dei prodotti inviati a Vqr proviene da soggetti scientificamente autorevoli e, allo stesso modo, risulta plausibile il verificarsi di una dinamica di segno opposto per cui vengano assegnati punteggi bassi a prodotti scientifici appartenenti ad autori con una "cattiva reputazione". Di questa possibile attribuzione di punteggi *biased* sono ben consapevoli i Gev, anche se tuttavia essi sostengono che «questo problema probabilmente si risolverà nel tempo dati i cambiamenti in corso nelle comunità di ricerca…» (Anvur, 2017a, p. 73-74). Probabilmente sarebbe stato possibile limitare tali effetti adottando una strategia di maggiore coinvolgimento e partecipazione, ad esempio, attraverso l'addestramento dei revisori all'uso pratico dello strumento. Inoltre, a conclusione dell'esercizio di valutazione, anche ai fini della progettazione delle future edizioni della Vqr, si sarebbero potuti analizzare gli esiti delle valutazioni alla luce dell'accoppiata tra le caratteristiche del

revisore e le caratteristiche del prodotto scientifico/autore del prodotto, classificando, ad esempio, per zona geografica dell'ateneo, per ruolo ricoperto, ecc. In altre parole, sarebbe stato di estrema rilevanza comprendere "chi ha valutato chi" in termini di caratteristiche generali[2].

4.3. Gli oggetti multipli e le conseguenze della sotto-determinazione semantica

Nel complesso, i primi tre quesiti della scheda sembrano ricalcare le definizioni dei criteri di qualità previsti nelle linee guida Anvur, adottandone perfino gli stessi termini. Le domande, formulate senza il ricorso a qualsiasi strategia di riduzione della complessità, presentano nella loro costruzione oggetti multipli, producendo una situazione di sotto determinazione semantica multipla (Fasanella e Martire, 2017, p. 98; cfr. Belleri, 2014; vedi anche Searle, 1979, 1980; Travis, 1975, 1981). Tale condizione induce ad accettare o a rifiutare in blocco tutti gli oggetti presenti nella domanda (Pitrone, 2009) con la conseguenza di non poter considerare comparabili le risposte fornite. I revisori si sono inevitabilmente trovati a esprimere un giudizio riferito ai diversi aspetti caratterizzanti i criteri di qualità attraverso l'assegnazione di un unico punteggio per ogni item.

Il criterio dell'originalità «ha lo scopo di misurare quanto sia innovativo il prodotto di ricerca, rispetto a un nuovo modo di pensare, nuove prospettive, tesi, temi e/o fonti, in relazione all'oggetto scientifico della ricerca, e pertanto quanto si distingua dai precedenti lavori sullo stesso tema» (Anvur 2017d, p. 2). Il testo della domanda corrisponde esattamente alla definizione che ne viene data in sede di bando ed è chiaramente sotto-determinato. Il quesito poteva essere scomposto ulteriormente in diverse sotto-domande, le quali a loro volta avrebbero dovuto essere sottoposte a processi di chiarificazione terminologica al fine di stabilire il più possibile un accordo intersoggettivo circa il loro significato.

Volendo scomporre il criterio dell'originalità per ogni oggetto presente nell'*item,* si potrebbero individuare le seguenti sotto-domande.

- Quanto è innovativo il prodotto di ricerca rispetto a un nuovo modo di pensare?

[2] L'argomento verrà sviluppato nel capitolo 6.

- Quanto è innovativo il prodotto di ricerca rispetto a nuove prospettive?
- Quanto è innovativo il prodotto di ricerca rispetto a nuove tesi?
- Quanto è innovativo il prodotto di ricerca rispetto a nuovi temi?
- Quanto è innovativo il prodotto di ricerca rispetto a nuove fonti?
- Quanto si distingue dai precedenti lavori sullo stesso tema?

A ben vedere, considerando singolarmente ognuno degli *item* così individuati, si potrebbero identificare alcuni casi di sinonimia, ad esempio un nuovo modo di pensare potrebbe essere considerato sovrapponibile con le nuove prospettive anche se, in ultima analisi, non possiamo averne certezza. In effetti, possiamo ipotizzare che nelle intenzioni dell'Anvur, ognuno dei termini inseriti nella definizione avrebbe dovuto essere portatore di un significato a sé stante ma, proprio in virtù di tale considerazione, ciò avrebbe richiesto un maggiore impegno nelle operazioni di disambiguazione terminologica proprio al fine di ridurre tale indeterminatezza.

Analogamente all'originalità, il rigore metodologico – che «ha lo scopo di misurare: i) il livello di chiarezza con cui il prodotto presenta gli obiettivi di ricerca; ii) il livello di competenza scientifica e di padronanza dello stato dell'arte; iii) la capacità di adottare una metodologia appropriata rispetto all'oggetto della ricerca; iv) il raggiungimento (realizzazione) degli obiettivi prefissati» (Anvur, 2017d, p. 2) – era composto da diverse sotto-dimensioni:

- livello di chiarezza con cui il prodotto presenta gli obiettivi di ricerca;
- livello di competenza scientifica;
- livello di padronanza dello stato dell'arte;
- capacità di adottare una metodologia appropriata rispetto all'oggetto della ricerca;
- raggiungimento/realizzazione degli obiettivi prefissati.

Nelle sotto-dimensioni sono presenti dei concetti complessi che dovrebbero, a loro volta, essere declinati attraverso dimensioni ancora più specifiche. In particolare, il concetto di competenza scientifica risulta estremamente astratto, tanto quanto quello di qualità della ricerca. Inoltre, la rilevazione delle competenze è sempre stato un terreno controverso e oggetto di accesi dibattiti nella comunità scientifica. Sostanzialmente, il livello di competenza scientifica si pone a un

grado di generalità più elevato rispetto agli altri oggetti presenti nell'*item*, risultando pertanto maggiormente esposto al rischio di essere compreso in modo diverso da ciascun revisore. Anche riguardo all'adozione di una metodologia appropriata valgono le considerazioni appena esposte circa il grado di generalità/astrazione, con la differenza che quest'ultima potrebbe addirittura identificarsi *in toto* con il criterio del rigore metodologico (cfr. capitolo 2). È necessario aggiungere che la messa a punto di tale sezione può essere considerata intrinsecamente e deliberatamente sotto-determinata vista la presenza di un elenco numerato all'interno della domanda formulata dal Gev 14.

Infine, per quanto riguarda l'impatto attestato o potenziale che «ha lo scopo di misurare il livello rispetto al quale il prodotto ha esercitato – o è presumibile eserciti in futuro – un'influenza sulla comunità scientifica presente, anche in base alla sua capacità di rispettare standard internazionali di qualità della ricerca» (Anvur, 2017d, p. 2), potevano essere individuate almeno tre sotto-domande:

- livello d'influenza esercitato dal prodotto sulla comunità scientifica;
- livello d'influenza che è presumibile il prodotto eserciti in futuro sulla comunità scientifica;
- capacità di rispettare standard internazionali di qualità della ricerca.

Riguardo quest'ultimo criterio, oltre al riferimento a un doppio arco temporale che contribuisce a rendere ancora più sotto determinata tale domanda, sussiste un problema che riguarda la possibilità di presumere e prevedere quale impatto avrà uno specifico prodotto sulla comunità scientifica. In altre parole, sarebbe possibile rilevare una potenzialità ancora inespressa e si ammette come giudizio valutativo un pre-giudizio, ossia ciò che si presume avverrà in futuro. Inoltre, in tal modo si sta implicitamente assumendo che tale criterio abbia una natura disposizionale. Pur volendo accogliere la fondatezza di tale assunto, risulterebbe estremamente rischioso tentare la rilevazione di caratteri disposizionali a meno di fare esplicito riferimento alle manifestazioni empiriche di tali disposizioni. Tuttavia, il caso della Vqr 2011-2014 non rientra nella situazione descritta.

Un ulteriore problema riguarda l'indicazione di giudicare il prodotto «anche in base alla sua capacità di rispettare standard internazio-

nali di qualità della ricerca», standard che, tuttavia, non vengono esplicitati, sollevando, in tal modo, una questione di interpretazione e di comparabilità delle valutazioni fornite dai revisori che, di volta in volta, potrebbero far riferimento a standard diversi. Vale la pena richiamare le argomentazioni già esposte nel capitolo 2 circa le diverse questioni che solleva una tale definizione del criterio dell'impatto. Ci si sta riferendo alla sovrapposizione tra un piano temporale (impatto passato/presente/futuro) e uno sostantivo (impatto teorico/applicativo) e al rapporto di dipendenza del criterio dell'impatto da standard internazionali di qualità della ricerca, oltre che alla possibilità che tale criterio possa, in realtà, essere considerato esterno alla qualità della ricerca. Inoltre, l'espressione "comunità scientifica presente" non è sufficientemente chiara ed è suscettibile di numerose interpretazioni differenti per diversi ordini di motivi.

In primo luogo, la comunità scientifica viene considerata come un insieme unitario e omogeneo, quando invece negli stessi documenti di rendicontazione redatti dai Gev e pubblicati dall'Anvur è possibile apprendere che si tratta di un insieme frammentato e in particolare alcune aree sono composte da «settori metodologicamente divisi in scuole, e privi di standard comuni e legittimati dall'insieme degli addetti» (Anvur, 2017a, p. 42). In secondo luogo, è utile precisare che la specificazione di cosa debba intendersi per "comunità scientifica presente" viene riportata esclusivamente nelle linee guida per Gev e revisori che sono state pubblicate in appendice al rapporto finale e, quindi, solo successivamente alla conclusione dell'esercizio di valutazione: «La comunità scientifica a cui riferirsi come contesto di comparazione (da utilizzare come contesto di riferimento) è quella presente e non l'intera storia delle idee dall'antichità ai giorni nostri» (Anvur, 2017c, 3). Peraltro, tale spiegazione non produce chiarimenti significativi visto che la definizione viene declinata in negativo, e cioè enunciando cosa non sia la "comunità scientifica presente" ma senza esplicitare cosa sia o comunque cosa si intenda.

In definitiva, volendo evitare la sotto-determinazione delle domande della scheda di valutazione, per ognuno dei tre criteri di qualità si potrebbe costruire una batteria di *item* a cui i revisori dovrebbero assegnare un punteggio secondo una scala Cantril. In tal modo, all'interno delle diverse batterie, una per ciascuno dei tre criteri, ogni oggetto presente nelle domande originarie troverebbe un suo corrispettivo

grazie ad un item appositamente predisposto. A questo punto non risulterebbe particolarmente problematica la ricomposizione dei punteggi ottenuti sugli item in un indice costruito per via matematica, considerando che i punteggi assegnati ai prodotti utilizzando una scala Cantril restituiscono variabili quasi-cardinali (Poli, 2008, p. 55) che si potrebbero trattare alla stregua di numeri naturali ricorrendo a procedure statistiche[3] (Pitrone e Pavisc, 2003, p. 134). Infine, ricomposti gli indici riferiti a ognuno dei tre criteri, potremmo procedere ad una loro ulteriore sintesi per via matematica o combinandoli tra loro mediante la riduzione di uno spazio di attributi, rispettando in tal modo anche l'assunto di ortogonalità implicito nella individuazione dei criteri costitutivi del concetto originario di qualità della ricerca operata da Anvur e Gev.

Deve peraltro essere segnalato che la predisposizione di una batteria di domande potrebbe incoraggiare la produzione di *response set* (Pitrone e Pavisc, 2003, p. 135). Tuttavia, potrebbe valer la pena correre questo rischio, in particolare, considerando l'elevata densità semantica riposta nei tre criteri di valutazione. In tal senso la proposta di utilizzare una sorta di sotto-criteri per la rilevazione della qualità dei prodotti scientifici, seppur rischi di rendere più corposa la scheda di valutazione, potrebbe produrre effetti positivi. Come si è avuto modo di osservare, i tre criteri presentano al loro interno alcune sotto-dimensioni che potrebbero risultare ridondanti o addirittura estranee rispetto al concetto originario. Seguendo principi di rilevanza e parsimonia si dovrebbe riuscire a prevenire una tale situazione, eliminando dalla batteria concetti-termini estranei o ridondanti, ottenendo uno strumento più agile e funzionale allo scopo.

4.4. La manipolazione *ex post* delle classi di merito: una questione di qualità

Dopo aver assegnato i punteggi per ognuno dei tre criteri, e prima di procedere alla formulazione di un giudizio esteso, ai revisori veniva mostrata la classe di merito finale (vedi capitolo 5) in cui il prodotto

[3] Tuttavia, nel trattare variabili quasi-cardinali è doveroso tenere sempre a mente che esse sono l'esito di un'operazione di *scaling* e non di misurazione (vedi Marradi, 2010).

 ISBN 9788835103219

ricadeva in base alla somma dei punteggi ottenuti sui tre criteri di qualità. Inoltre, i revisori dovevano confermare l'assegnazione alla classe di merito oppure modificare i punteggi precedentemente conferiti e confermare la nuova collocazione del prodotto in una delle altre classi di merito. Questo iter, già emerso come problematico per la Vqr 2004-2010 (Fasanella e Di Benedetto, 2014), era finalizzato a intervenire «nel caso in cui la classe di merito proposta non corrisponda alla percezione generale della qualità del prodotto valutato» (Anvur, 2017c, p. 5). Tuttavia, la procedura rende possibile la manipolazione *ex post* dei punteggi assegnati ai tre criteri al fine di collocare il prodotto in una classe di merito indipendentemente dai punteggi conferiti nella prima valutazione, indebolendo la solidità dei giudizi dei revisori in termini di stabilità e veridicità e senza fornire evidenze della maggiore affidabilità della seconda valutazione espressa. Peraltro, non è chiaro se venga conservata traccia delle modifiche dei punteggi e della classe di merito. Se l'intenzione era quella di rilevare la percezione generale della qualità del prodotto indipendentemente dal giudizio espresso sui tre criteri, sarebbe stato opportuno predisporre una sezione dedicata esclusivamente a tale fine, evitando di produrre un errore indotto (Fasanella e Martire, 2017, p. 99-100) dalla struttura stessa della scheda di valutazione. Va da sé che la possibilità di cambiare i punteggi rispetto ai tre criteri in funzione di una riallocazione della classe di merito introduce ulteriori interrogativi rispetto alle soglie che l'Anvur ha stabilito per rientrare nelle stesse. Ciò, infatti, potrebbe indicare che il significato veicolato dalle classi di merito (cfr. capitolo 5) – Eccellente, Elevata, Discreta, Accettabile, Limitata – non sia il medesimo per *referees*, Gev e Anvur e di conseguenza, per ciascuno di essi, il punteggio soglia necessario per rientrare in ognuna delle classi possa essere differente e non condiviso. D'altra parte, possiamo ipotizzare che nel giudizio complessivo di qualità, per i revisori siano intervenuti criteri altri (cfr. capitolo 2) rispetto a quelli individuati dall'Anvur e riproposti dai Gev per rappresentare il concetto di qualità della ricerca. In quest'ultimo caso, sarebbe fondamentale identificare quali standard siano intervenuti nella seconda valutazione per la riallocazione della classe di merito in vista di un processo di valutazione che sia in grado di assicurare trasparenza delle procedure e possibilità di apprendimento. A tale proposito, la predisposizione di una sezione per la rile-

 ISBN 9788835103219

vazione della percezione generale della qualità del prodotto può consentire – se corredata di precise istruzioni – l'esplicitazione dei criteri intervenuti nella determinazione del giudizio di qualità, ma potrebbe anche fornire indicazioni utili a una definizione del concetto di qualità condivisa da comunità scientifiche e apparato istituzionale (capitolo 6).

Il tentativo di standardizzare un concetto complesso, quale quello della qualità della ricerca, è stata un'operazione articolata il cui risultato finale è dipeso da alcune decisioni fondamentali. Il raggiungimento di tale obiettivo avrebbe richiesto un approccio metodologico chiaramente definito: se da una parte era possibile intraprendere la via della costruzione di uno strumento di rilevazione massimamente standardizzato, dall'altra era doveroso chiedersi se l'uniformazione degli stimoli che costituiscono la scheda di valutazione fosse una condizione sufficiente a garantire la piena comparabilità delle risposte fornite o se, invece, sarebbe stato necessario tentare la via della standardizzazione del significato, ossia la messa in atto di operazioni di specificazione dei significati che non sottovaluti i processi di interpretazione sottesi alle operazioni che vanno dalla formulazione di una domanda alla registrazione della risposta (Mauceri, 2003, p. 90-91), tanto più considerando che uno stesso termine può assumere molteplici significati e che termini differenti possono assumere uno stesso significato anche per una medesima persona in momenti diversi[4], quindi sia in termini intersoggettivi che in termini intrasoggettivi. In altre parole, «poiché l'attenzione di comunità scientifiche differenti è incentrata su campi di indagine differenti, la comunicazione professionale attraverso i confini che separano un gruppo da un altro è talvolta molto difficile, spesso dà luogo a fraintendimenti, e può, se cercata, suscitare un disaccordo significativo e precedente non sospettato» (Kuhn, 1969; trad. it. 2009, 214-215). Sarebbe stato auspicabile, dunque, tentare una standardizzazione fondata sulla congruenza di significato.

A tale proposito, un aspetto complementare riguarda la cosiddetta stabilità dello strumento «ossia la capacità dello strumento osservativo di dar luogo a risultati pressoché analoghi al variare dell'unità di tempo in cui lo stesso fenomeno viene rilevato» (Mauceri, 2003, p. 56). Nello specifico caso della Vqr la stabilità del giudizio dei revisori si pone come un requisito fondamentale per lo svolgimento della *peer*

[4] Per una trattazione completa si rimanda a Marradi (1994).

review, la quale «si basa sul presupposto essenziale che i revisori condividano una serie di conoscenze e valori su cui fondare la valutazione, e che dunque i loro pareri siano basati ed espressi lungo un'unica scala di giudizio» (Fasanella e Di Benedetto, 2015, p. 66). In sintesi, le scale di giudizio dovrebbero poter essere considerate stabili sia al livello di inter-revisore e quindi tra diversi revisori, sia al livello intra-revisore per cui il medesimo revisore si riferisce alla stessa scala in momenti differenti, ma anche nel valutare prodotti scientifici distinti. In effetti, già dalle pagine introduttive del rapporto di Area 14, emerge che «in un'area eterogenea come quella presa in analisi dal Gev 14 manca ancora una metodologia comune di valutazione. Il Gev ha cercato, quindi, di identificare alcuni standard che permettessero di valutare tutti i prodotti di ricerca presentati, standard che fossero allo stesso tempo selettivi ed inclusivi. (…) Anche se nell'area di riferimento non si dispone di principi, metodi e criteri condivisi, tuttavia una certa convergenza si sta determinando almeno su alcuni standard» (Anvur, 2017a, pp. 14-15). E ancora, «L'esiguità di tali Ssd – soprattutto quelli che hanno per oggetto i cosiddetti "studi di area" – insieme all'interpretazione talora molto diversa che gli afferenti hanno della propria disciplina rendono assai delicato e complesso il processo di valutazione tra pari» (*ivi* p. 20-21). Quali siano gli standard sui quali si è determinato un certo grado di accordo e il perché non vengano esplicitati in sede di rendicontazione resta una questione irrisolta che necessiterebbe un approfondimento da parte dei Gev. Seppure dagli estratti del rapporto Anvur sia possibile comprendere che si starebbe determinando una convergenza su alcuni standard possiamo affermare che quest'ultima sarebbe ancora in una fase di incubazione. Va detto, inoltre, che le affermazioni sopra riportate – oltre a mettere in luce come la stabilità del giudizio non possa essere garantita – mettono in discussione alcuni principali assunti della revisione tra pari poiché, almeno per l'Area 14, non si dispone di principi, metodi e criteri condivisi.

Altro presupposto su cui si basa l'esercizio di valutazione *peer* è la sincerità del giudizio dei revisori. A tale proposito, la Vqr prevedeva l'assunzione di un codice etico da parte dei revisori e dei membri del Gev, i quali si sono impegnati a rispettarlo e a uniformarsi ad esso, garantendo il rispetto dei principi di imparzialità, lealtà alla comunità scientifica e riservatezza. Tuttavia, è opportuno tenere in considera-

zione che l'esercizio Vqr si caratterizza come *single-blind*; ciò significa che l'autore del prodotto sottoposto a valutazione non è anonimo: «Nella valutazione *double-blind*, imparzialità e riservatezza sono iscritti nelle stesse circostanze del giudizio.

Invece nel caso della Vqr, dove non c'è anonimato del valutato, sta al revisore esercitare uno sforzo di cecità per garantire l'imparzialità, in primis, e la riservatezza poi» (Anvur, 2017b, p. 11). Eppure, assieme al testo del prodotto in formato pdf, al revisore sono state fornite le informazioni ad esso relative, contenute nella scheda prodotto compilata dall'autore, come gli indici bibliometrici, se presenti, la classificazione della rivista, nel caso di articolo, e il superamento di una valutazione *peer review* in caso di monografia o saggio in volume (*ivi*, p. 9). Nella maggior parte dei casi di valutazione tra pari l'identità del revisore è nascosta (*single-blind*) per incoraggiare i commenti genuini proteggendoli da possibili rappresaglie di autori (Lee *et al*., 2013), questi ultimi, dal canto loro, rimangono esposti alle posizioni soggettive tipiche della *peer review,* posizioni che sono spesso assai distanti nel campo delle scienze sociali non solo a causa dell'appartenenza a scuole e approcci teorici e/o metodologici differenti, ma anche per questioni di conflittualità accademiche che spesso hanno poco a che fare con il tema della qualità dei prodotti scientifici. In tal senso, appare inconciliabile la scelta di fornire ai revisori le informazioni contenute nella scheda del prodotto con l'aspettativa che tali informazioni vengano considerate in un contesto di *informed peer review*, senza predeterminare la valutazione (Anvur, 2017b, p. 11). Se da una parte viene richiesto ai revisori di mettere tra parentesi *bias* di carattere teorico (*ibidem*), dall'altra sembrerebbe non siano stati tenuti debitamente in considerazione gli effetti distorsivi largamente documentati nella letteratura in tema di revisione tra pari (Xie, 2014; Lee *et al*., 2013; Wennerâs e Wold, 1997; Mahoney, 1977) e, soprattutto, già emersi con riferimento alla Vqr 2004-2010 (Fasanella e Di Benedetto, 2015).

L'effetto San Matteo (Merton, 1968a; 1968b) opera in rapporto al prestigio e al ruolo accademico aumentando i riconoscimenti scientifici e la possibilità di pubblicare per chi gode già di una buona reputazione, mentre li riduce per coloro che non ne godono. Anche prendendo in considerazione la produttività delle istituzioni scientifiche è stato segnalato un simile effetto definito vantaggio cumulativo istitu-

zionale (Fasanella e Di Benedetto, 2015, p. 54-55; Bentley e Blackburn, 1990; Merton, 1968a; Crane, 1967).

Notoriamente nella *peer review* è riscontrabile il cosiddetto effetto alone per cui ricoprire una posizione in un'università prestigiosa «pone una sorta di "aureola" sul lavoro di un uomo, in modo che appaia meglio ai suoi colleghi di quanto potrebbe altrimenti» (Crane, 1965, p. 700).

Malgrado tali *bias* non possano essere eliminati completamente, un'analisi delle caratteristiche dei revisori, dei prodotti e degli autori potrebbe consentire un loro maggiore controllo e ridurre le possibilità di essere soggetti a tali distorsioni attraverso un'eventuale eliminazione dalla lista di quei revisori che, invece, ne risultano maggiormente affetti (cfr. capitolo 6).

Rispetto, invece, al tentativo di superare le criticità connesse al problema della manipolazione *ex post*, si potrebbe pensare di far compilare la scheda di valutazione assegnando i punteggi sui tre criteri, per chiedere successivamente ai *referees* di inviare conferma dei punteggi. Solo a questo punto sarebbe possibile visualizzare la classe di merito finale nella quale il prodotto valutato è stato collocato, abrogando la possibilità di modificare i punteggi assegnati ai criteri. I revisori, grazie alla presenza della sezione dedicata al giudizio generale di qualità, potrebbero poi esprimere l'accordo o il disaccordo sulla classe di merito finale nella quale è stato ricondotto il prodotto, proporre quella in cui intendevano collocarlo ed esplicitare quali criteri siano intervenuti. Una simile procedura produrrebbe materiale empirico molto utile per studiare la complessità semantica del concetto di qualità della ricerca e delle sue dimensioni e per la sua applicabilità agli esercizi di valutazione[5].

Accanto a tali possibili correttivi, sarebbe necessario un periodo di addestramento dei revisori. Simili procedure, d'altronde, fanno parte delle ordinarie operazioni di qualsiasi ricerca o esercizio di valutazione che si basi sul coinvolgimento di diversi attori e inoltre potremmo considerarla anche coerente con il proposito di una valutazione democratica e partecipata (Stame, 2016). Un training dei revisori finaliz-

[5] Nel capitolo 6 verrà esposta più approfonditamente la proposta di ricerca che potrebbe svilupparsi a partire dalle informazioni fornite da una simile procedura.

zato a potenziare le competenze valutative potrebbe dimostrarsi vantaggioso anche per ridurre gli errori nelle valutazioni e accrescere la qualità della revisione (Fasanella e Di Benedetto, 2015, p.56; Schroter *et al*., 2004). D'altra parte, potremmo ipotizzare l'esistenza di una sorta di propensione al giudizio che potrebbe scaturire da alcune caratteristiche possedute dai revisori. A titolo esemplificativo, la propensione dei *referees* a fornire giudizi valutativi potrebbe dipendere dalla relazione tra ruolo ricoperto dai revisori e ruolo ricoperto dall'autore del prodotto – professore ordinario, associato, ricercatore –, dalla loro affiliazione, ecc. (cfr. capitolo 6)

Va aggiunto che sarebbe stato utile svolgere interviste cognitive (Liani e Martire, 2017; Tusini, 2006) ad un campione di aspiranti revisori opportunamente scelto. Tale tecnica di *pretest* permetterebbe di collaudare sia lo strumento di rilevazione del giudizio, sia la correttezza del processo di operativizzazione. Anche se una simile strategia doveva essere predisposta in sede di progettazione dello strumento di rilevazione e quindi *ex ante* tuttavia, attraverso l'analisi del materiale empirico, in particolare l'esame dei giudizi estesi, potevano essere predisposti una serie di controlli *ex post*.

4.5. I giudizi obbligatori: un'occasione mancata

Come già anticipato, l'ultima sezione della scheda di valutazione dei prodotti è stata dedicata a un campo libero per la formulazione di un giudizio sintetico finale che potesse fornire una motivazione dei punteggi attribuiti al prodotto attraverso i tre criteri di qualità. A tale campo, reso obbligatorio a partire dall'esercizio valutativo del 2011-2014, è stata assegnata una funzione di *accountability* del processo di valutazione e di risoluzione in caso di giudizi difformi su uno stesso prodotto. In effetti, nel rapporto finale di Area 14 è possibile leggere che «… a differenza della precedente Vqr, in quest'occasione i revisori erano tenuti a giustificare il punteggio assegnato al prodotto con una motivazione. Questa richiesta, oltre a rendere più trasparente il processo di valutazione, è risultata molto utile nella fase di convalida delle valutazioni in caso di giudizi difformi» (Anvur, 2017a, p. 23). Inoltre, dai documenti pubblicati a conclusione dell'esercizio di valutazione è possibile apprendere che «Nel campo libero contenente le motivazioni

dell'attribuzione della classe di valutazione al prodotto il revisore è invitato a illustrare in quale misura abbia tenuto conto delle informazioni della scheda prodotto» (Anvur, 2017b, p. 9).

Tuttavia, stando al rapporto finale dell'Anvur e ai vari documenti disponibili non vi è traccia di tali giudizi. Anche se da questo estratto è possibile apprendere che tale campo è stato tendenzialmente compilato apportando le motivazioni dei punteggi assegnati, risulta però arduo affermare che esse abbiano reso più trasparente il processo di valutazione, considerando che tali motivazioni, al momento, non sono in alcun modo reperibili o consultabili. In effetti, nel rapporto finale di Area 14 è possibile leggere: «…è difficile capire se la ragione di questa parsimonia nelle valutazioni alte, particolarmente evidente negli Ssd di sociologia, dipenda dalla severità dei valutatori o dalla qualità stessa della ricerca; non disponendo di criteri oggettivi, quali quelli bibliometrici, la valutazione *peer* potrebbe essere *biased*, soprattutto in settori metodologicamente divisi in scuole, e privi di standard comuni e legittimati dall'insieme degli addetti»[6] (Anvur, 2017a, p. 42). In realtà, il giudizio sintetico obbligatorio avrebbe dovuto ricoprire un ruolo dirimente nel rendere maggiormente ripercorribile il processo di valutazione, soprattutto nella fase di convalida delle valutazioni in caso di giudizi difformi. Al contrario, si potrebbe affermare che tale campo abbia aperto le porte a ulteriori elementi di discrezionalità nelle valutazioni e anche a processi difficilmente controllabili visto che «nel caso in cui si sia notato una chiara discrepanza fra voto e giudizio, o la mancata considerazione dei dati concernenti la sede di pubblicazione, o una netta distanza tra i due giudizi, i membri del Gev sono intervenuti proponendo interventi correttivi» (Anvur, 2017a, p. 71), interventi correttivi non meglio specificati che potrebbero inficiare e svilire il lavoro svolto dai *referees*.

Inoltre, l'assenza totale di indicazioni relative alla compilazione di tale campo obbligatorio rende assolutamente non comparabili e quindi anche poco utilizzabili i giudizi espressi. Per tale motivo si rende necessario fornire istruzioni dettagliate sull'utilizzo del campo libero, so-

[6] L'oggettività, in questo caso, sembrerebbe far riferimento ad una capacità intrinseca degli indicatori bibliometrici di rappresentare validamente la qualità della ricerca senza alcun rischio di distorsione. Tuttavia, ad oggi è in corso un ampio dibattito sui limiti e i vantaggi del ricorso ad indicatori bibliometrici (nell'ambito delle scienze sociali vedi Bonaccorsi, 2012).

prattutto in considerazione della fondamentale funzione ad esso assegnata in sede di rapporto finale. In effetti, questa considerazione può valere anche per la scheda nel suo insieme: essa è infatti calata in un vuoto procedurale in cui risultano assenti tutti gli strumenti necessari ad una sua agevole applicazione. In sostanza, non potendo essere data per scontata la competenza dei *referees* a fornire un giudizio valutativo attraverso un qualsiasi strumento, si rende necessaria, per tali motivi, la predisposizione di linee guida all'uso pratico della scheda, linee guida che devono necessariamente pervenire ai revisori in anticipo rispetto all'utilizzo della scheda stessa.

I giudizi obbligatori avrebbero dovuto ricoprire un ruolo strategico nel più generale processo di valutazione, anche rispetto ad alcune fasi maggiormente problematiche come, ad esempio, la possibilità di cambiare i punteggi. Se dovutamente strutturata, l'espressione verbale del giudizio avrebbe non solo potuto far emergere i significati conferiti alle classi di merito, ma anche raccogliere le motivazioni sottostanti all'alterazione dei punteggi in funzione della riallocazione finale del prodotto. Anche rispetto all'assunto, non ancora dimostrato, che i tre criteri abbiano lo stesso peso nel contribuire alla formazione del concetto di qualità della ricerca i giudizi estesi potrebbero fornire elementi utili alla ponderazione dei criteri che, ad esempio, potrebbero essere ordinati secondo la loro importanza. In tale maniera «Il punteggio assegnato al prodotto su ciascun criterio potrebbe assumere un peso proporzionale alla centralità del criterio stesso rispetto alla determinazione della qualità della ricerca» (Fasanella e di Benedetto, 2015, p. 65). Per questa via, un aspetto su cui vale la pena riflettere riguarda la tipologia di prodotti che viene sottoposta a revisione. In effetti, estendendo il ragionamento esposto in precedenza circa l'attribuzione di una ponderazione dei punteggi ottenuti applicando i tre criteri di qualità, possiamo debitamente ritenere che sia possibile applicare ponderazioni differenti in funzione dei tipi di prodotto di volta in volta presi in esame. Tale considerazione nasce dal fatto che alcuni tipi di prodotto hanno ottenuto valutazioni sbilanciate nelle classi di merito – che derivano dalla ricomposizione in un indice sintetico dei punteggi attribuiti ai tre criteri di qualità. Ad esempio, nel rapporto di Area 14 si può leggere che «i contributi in volume sono chiaramente il tipo di prodotto che ottiene meno valutazioni nelle due classi di merito più alte e più valutazioni nelle ultime due» (Anvur, 2017a, p. 44).

 ISBN 9788835103219

Tali ipotesi di lavoro dovrebbero essere approfondite maggiormente sulla base di evidenze empiriche, in parte possibile con i dati a disposizione dell'Agenzia, se si considerasse la possibilità di un loro utilizzo a fini analitico-conoscitivi. Resta, comunque, il dubbio che l'agenzia abbia la possibilità di visionare i punteggi sui singoli criteri, considerando che la scheda così come progettata non sembrerebbe conservare traccia di tutti i passaggi precedenti all'assegnazione della classe di merito finale e che i giudizi estesi non sono ancora stati resi consultabili. Tutte queste informazioni sarebbero state di estrema utilità non solo ai fini della comprensione degli esiti della Vqr 2011-2014, ma anche per un futuro esercizio di valutazione della qualità della ricerca[7]. Si tratta in sostanza, di una vera e propria occasione mancata.

[7] Si rimanda al capitolo 6 per l'individuazione delle linee di ricerca basate sulla consultazione dei giudizi estesi.

 ISBN 9788835103219

5. Formulazione e sintesi dei giudizi

di Federica Fusillo

5.1. La classificazione della qualità della ricerca

Tra gli obiettivi finali della Vqr vi era il conferimento di una classe di merito, tra le cinque individuate, a ciascun prodotto (Anvur, 2017c, p.4). La compilazione della scheda di valutazione era quindi propedeutica all'assegnazione di un punteggio, ottenuto dalla somma delle valutazioni che ogni revisore doveva fornire sui tre criteri. Tale giudizio di qualità era finalizzato all'attribuzione dei prodotti a una delle classi di merito, definite sulla base di soglie stabilite a monte, dei punteggi dei singoli revisori.

Una qualsiasi classificazione di un concetto sottintende operazioni di astrazione e sintesi, risponde agli obiettivi di ordine, semplificazione e organizzazione ed è dotata di natura negoziale e validità contestuale; per questi motivi, uno schema di classificazione può essere definito giusto e utile, ma non certamente vero (Faggiano, 2012, p. 24). Tuttavia, pur essendo l'esito di una stipulazione culturalmente mediata, non è possibile immaginare un sistema di classificazione privo di regole metodologiche che lo giustifichino.

Infatti, al fine di giungere a una corretta suddivisione dell'estensione di un concetto, è necessario rispettare alcune condizioni. Innanzitutto, il numero di classi deve essere esaustivo, in modo tale che nessun oggetto/soggetto rimanga escluso[1].

[1] A tal proposito, già Lazarsfeld (1951; tr. it., 1961, pp. 239) proponeva di aggiungere, in casi estremi, una categoria residuale, nella quale ricondurre tutti i soggetti che non potessero rientrare nelle altre classi; tuttavia, il metodologo austriaco, e successivamente la letteratura sul tema (Marradi, 1984, p. 76), si riferiva principalmente al problema della categorizzazione

 ISBN 9788835103219

La seconda condizione da rispettare per una buona classificazione è che non vi sia alcun tipo di sovrapposizione tra due o più classi e, di conseguenza, che ogni oggetto/soggetto rientri in una e una sola di esse (mutua esclusività). Infine, il criterio in base al quale viene operata la suddivisione (*fundamentum divisionis*) deve essere unico. Per quanto riguarda l'ultima delle tre condizioni, è bene precisare che nel caso in cui i criteri di suddivisione fossero più di uno, allora non si parlerebbe di una semplice classificazione del concetto, ma di una tipologia, e non di classi, ma di tipi (Lazarsfeld e Barton, 1951; tr. it., 1961, p. 239-240; Marradi, 1984, p. 74; Campelli, 2001, pp. 15-16; Corbetta, 2014, p. 520).

Osservando la definizione lessicale che è stata fornita delle cinque classi di merito per la Vqr 2011-2014 (d.m. n. 458 del 2015, art. 5; Anvur, 2015, 14-15), riportata di seguito, è proprio questo il caso: la suddivisione in classi ordinate del concetto di qualità, essendosi basata su tre criteri, ha restituito più che delle categorie, dei tipi:

- *Eccellente*: la pubblicazione raggiunge i massimi livelli in termini di originalità e rigore metodologico, e ha conseguito o è presumibile che consegue un forte impatto nella comunità scientifica di riferimento a livello internazionale.
- *Elevato*: la pubblicazione raggiunge buoni livelli in termini di originalità e rigore metodologico, e ha conseguito o è presumibile che consegua un impatto significativo nella comunità scientifica di riferimento a livello internazionale e/o nazionale.
- *Discreto*: la pubblicazione raggiunge discreti livelli in termini di originalità e rigore metodologico, e ha conseguito o è presumibile che consegua un apprezzabile impatto nella comunità scientifica di riferimento a livello internazionale e/o nazionale.
- *Accettabile*: la pubblicazione raggiunge livelli sufficienti in termini di originalità e rigore metodologico, e ha conseguito o è presumibile che consegua un impatto circoscritto nella comunità scientifica di riferimento a livello internazionale e/o nazionale.

delle risposte fornite da intervistati a domande aperte. Nell'ambito di ricerca che si sta trattando in questa sede, non sembrano esserci le condizioni per prevedere casi simili, ossia che non rientrino in nessuna delle classi di merito previste. Anzi, una pubblicazione che fosse risultata non valutabile dal punto di vista dei tre criteri di qualità non sarebbe stata assolutamente auspicabile, poiché avrebbe rappresentato o rappresenterebbe una spia sufficientemente importante di una cattiva e frettolosa concettualizzazione del concetto da rilevare.

 ISBN 9788835103219

- *Limitato*: la pubblicazione raggiunge un livello scarso di originalità e rigore metodologico e ha conseguito o è presumibile che consegua un impatto molto limitato nella comunità scientifica di riferimento a livello internazionale e/o nazionale.

Tuttavia, esse descrivono solamente dei tipi ideali, in cui le valutazioni fornite sui tre criteri risultino perfettamente concordanti, perciò non sembrano aver rispettato la prima condizione di una corretta classificazione, ossia l'esaustività (Di Benedetto, 2015, p. 107). Nel bando, tuttavia, in seguito alla definizione delle classi, si può leggere: «le definizioni dei livelli di qualità hanno carattere esemplificativo in quanto fanno riferimento esclusivamente ai casi in cui le valutazioni attribuite ai tre criteri risultino concordanti» (Anvur, 2015, p. 15).

La consapevolezza di tale lacuna, evidente nella specifica appena riportata, sembra un espediente per sollevare l'agenzia e il Miur dalla responsabilità di dover definire i casi discordanti, lasciando ai Gev la piena libertà di stabilire le regole di una loro classificazione e le soglie da applicare ai punteggi finali per la definizione del livello di qualità dei prodotti. Tale mancanza, pur essendo plausibile da un punto di vista logico, considerando le possibili difficoltà riscontrabili nel tentativo di spalmare un unico schema concettuale a tutte le aree disciplinari, è possibile che abbia aggravato, da un punto di vista operativo, i compiti dei Gev, i quali si sono trovati così a dover stabilire, senza alcuna indicazione, le regole per la classificazione dei prodotti.

Rimandando la discussione sulle soglie stabilite e sui casi di giudizi discordanti tra i due revisori ai paragrafi che seguono, è il caso di effettuare un confronto tra le definizioni lessicali proposte per la Vqr 2011-2014 e quelle del precedente esercizio. Nella Vqr 2004-2010 lo schema classificatorio prevedeva quattro livelli, come segue (Anvur, 2011, pp. 7-8):

- i prodotti di livello *eccellente* sono quelli riconosciuti come eccellenti a livello internazionale per originalità, rigore metodologico e rilevanza interpretativa; oppure quelli che hanno rinnovato in maniera significativa il campo degli studi a livello nazionale;
- i prodotti di livello *buono* sono quelli d'importanza internazionale e nazionale riconosciute per originalità dei risultati e rigore metodologico;

 ISBN 9788835103219

- i prodotti di livello *accettabile* sono quelli a diffusione internazionale o nazionale che hanno accresciuto in qualche misura il patrimonio delle conoscenze nei settori di pertinenza;
- i prodotti di livello *limitato* sono quelli a diffusione nazionale o locale, oppure in sede internazionale di non particolare rilevanza, che hanno dato un contributo modesto alle conoscenze nei settori di pertinenza.

Saltano subito all'attenzione alcune particolari differenze: innanzitutto, si è passati da quattro classi a cinque, con la divisione della precedente classe "Buono" in "Elevato" e "Discreto". È probabile che l'intenzione fosse quella d'isolare maggiormente i prodotti di profilo alto dal resto dei contributi inviati a valutazione.

C'è da dire, inoltre, che ad una maggiore articolazione di uno schema di classificazione corrisponde, normalmente, una maggiore omogeneità delle classi. Come giustamente osserva Lazarsfeld, ogni suddivisione di un concetto tende a semplificare e sintetizzare innumerevoli aspetti di un fenomeno, con il rischio, nel caso in cui la classificazione risultasse estremamente semplice, di perdere di vista importanti distinzioni. D'altra parte, è pur vero che prevedere un numero eccessivo di raggruppamenti può rendere la successiva elaborazione dei dati particolarmente ostile (Lazarsfeld e Barton, 1951; tr. it., 1961, p. 239-240; Marradi, 1984, pp. 74-75). Stante ciò, senza alcuna pretesa di voler proporre un numero di classi di merito ideale, è senza dubbio apprezzabile la volontà di aumentare la sensibilità dello schema classificatorio per la Vqr 2011-2014.

Ulteriore particolare degno di nota è rappresentato dalla definizione delle classi di merito maggiormente articolata e costruita sulle tre dimensioni della qualità della ricerca, le quali vengono richiamate continuamente. Si ricorda che nella Vqr 2004-2010 essa risultava poco chiara ed esaustiva e non si basava interamente sulle dimensioni del concetto di qualità della ricerca, assumendo, da una parte, un certo legame tra i criteri della rilevanza e dell'originalità e, dall'altra, attribuendo implicitamente un ruolo poco decisivo al terzo criterio dell'internazionalizzazione (Fasanella e Di Benedetto, 2014, p. 78). Sembrerebbe, quindi, che ci sia stata maggiore attenzione nella determinazione dei livelli di qualità nei quali i prodotti possono ricadere.

Tuttavia, il riferimento al livello di originalità, rigore metodologico

e impatto attestato o potenziale che il prodotto doveva raggiungere per rientrare in una determinata classe sembra ancora connotato da una certa vaghezza, determinata in buona parte da una definizione operativa del concetto definita nei capitoli precedenti "tronca". Gli aggettivi utilizzati per definire i livelli di qualità ("massimo", "buono", "discreto", "accettabile" e "limitato" per l'originalità e il rigore metodologico, e "forte", "significativo", "apprezzabile", "circoscritto" e "molto limitato" per l'impatto) risultano ancora troppo generici e poco esplicativi della dimensione a cui si riferiscono (d.m. n. 458 del 2015, art. 5; Anvur, 2015, pp. 14-15). Tale vaghezza semantica potrebbe essere compensata, in parte, da un adattamento dello schema classificatorio allo schema mentale della comunità scientifica oltre che dalla logica ordinale sottintesa alle stesse classi di merito.

Sarebbe stato auspicabile, in particolar modo nell'area delle scienze politiche e sociali, un qualche sforzo per specificare e disambiguare eventualmente alcuni termini delle definizioni lessicali, come avvenuto in altre due aree in particolare. Il Gev 12, per esempio, oltre ad aver operato un importante lavoro sulla specificazione del significato della dimensione della qualità della ricerca in un'ottica di stampo prettamente giuridico, ha anche lavorato in vista di una definizione lessicale maggiormente accurata delle classi di merito. In particolare, sono stati aggiunti alcuni termini per precisare il livello d'impatto potenziale o effettivo che un prodotto doveva raggiungere per rientrare nelle tre classi di merito superiori. In questo modo, un contributo doveva fungere da punto di riferimento "di prim'ordine", "importante" o "utile" per lo studio del tema, per poter rientrare, rispettivamente, nelle classi "eccellente", "elevato" o "discreto" (Anvur, 2017, pp. 4-5). Un analogo lavoro è stato effettuato dagli Esperti Valutatori dell'Area 11b. Più che lavorare sulle definizioni lessicali, tuttavia, sono state inserite alcune limitazioni per l'assegnazione di determinate classi di merito a specifici tipi di prodotti; ed è così che una monografia poteva ambire all'eccellenza solamente se a diffusione internazionale, mentre un contributo in volume poteva rientrare al massimo nella classe "discreto", se pubblicato su volume internazionale, o "accettabile", se in volume nazionale. Le condizioni erano così vincolanti che, nel caso in cui un revisore avesse fornito un giudizio superiore, il sistema avrebbe restituito automaticamente un errore e la scheda non sarebbe stata inviata finché il voto finale non fosse stato modificato (Anvur, 2017i, p. 5).

5.2. La sintesi del giudizio del singolo revisore

Le definizioni lessicali delle classi di merito avevano, implicitamente, lo scopo di guidare i revisori nella fase operativa di espressione di un giudizio sui prodotti. Idealmente, avendo visionato precedentemente i livelli di qualità che un prodotto doveva raggiungere per rientrare in una determinata categoria, il revisore doveva esprimersi secondo i tre criteri, assegnando in base a ciascuno di essi un punteggio che rispecchiasse il proprio parere in merito; la somma di tali voti, confrontata con i valori soglia prestabiliti dall'Anvur, permetteva di assegnare il contributo direttamente ad una delle classi di merito discusse precedentemente.

Al di là della discutibile possibilità concessa ai revisori di modificare il proprio giudizio *ex ante*, potendo visualizzare a margine della scheda quale classe di merito fosse associata ai punteggi assegnati[2], le condizioni essenziali, elencate da Marradi (2007, pp. 187-189), che un indice sommatorio deve soddisfare sembrerebbero in questo caso esser state rispettate:

- non erano presenti dati mancanti, in quanto i revisori erano tenuti a fornire un giudizio su tutti e tre i criteri, senza la possibilità di escluderne qualcuno;
- le scale dei punteggi avevano la stessa estensione e "direzione";
- non vi era alcuna ragione valida per procedere alla ponderazione dei criteri prima di effettuare la sommatoria finale, poiché ognuno di essi contribuiva allo stesso modo alla definizione del concetto originario.

Una volta compilata la scheda e definito il punteggio finale, esso veniva automaticamente confrontato con le soglie prestabilite e, infine, il prodotto valutato ricondotto alla classe di merito di pertinenza. Si è trattato, pertanto, di utilizzare un semplice espediente matematico (la somma dei punteggi), al fine di giungere rapidamente ad una classificazione realizzata con tre *fundamenta divisionis*, ossia ad una tipologia (Nobile, 2008, p. 70). Le condizioni da rispettare per individuare

[2] Si rimanda al capitolo 4 sull'analisi della scheda di valutazione per tutti i problemi legati all'utilizzo di una scala per la valutazione della qualità dei prodotti.

quali e quante categorie debba possedere l'indice finale sono pressoché le stesse che guidano la concettualizzazione di una classificazione: i casi inclusi nella stessa categoria devono essere massimamente omogenei, mentre i casi inclusi in categorie diverse devono essere massimamente eterogenei; inoltre, il punto di riferimento, in questa fase, non devono più essere le singole dimensioni, bensì il concetto generale di qualità della ricerca (Marradi, 1984, p. 115).

È questo il momento in cui, attraverso un espediente tecnico, si ritorna al concetto originariamente definito operativamente, sintetizzando il suo significato in un unico valore. Fase particolarmente delicata, per la quale già Lazarsfeld aveva previsto diverse strade da poter percorrere, in base alle scelte effettuate nella definizione operativa del concetto originario. Infatti, è possibile ritornare ad esso attraverso diverse operazioni di combinazione delle informazioni ottenute sui singoli indicatori (per somma, per rapporto, per differenza, per via tipologica, ecc.), facendo particolare attenzione ad alcuni criteri logico-procedurali, come (Agnoli, 1994, p. 166):

- la funzione che i singoli indicatori hanno in relazione al concetto originario;
- il rapporto che gli indicatori hanno tra di loro, in relazione alla copertura semantica del concetto originario;
- i tipi di operazioni che si possono effettuare con le variabili ottenute.

Il fatto che nel passaggio dai valori numerici alle classi non siano stati presi alcuni accorgimenti logici, primo fra tutti una divisione equilibrata della scala di punteggi, ha portato a una classificazione poco chiara e non immediatamente comprensibile, quando la procedura è finalizzata, per sua natura, a mettere in ordine, distinguere e dotare di significato una realtà più complessa (Faggiano, 2012, p. 12).

Osservando la Tab. 1, che illustra la corrispondenza tra i punteggi e le classi di merito nel caso della valutazione di un revisore, è possibile notare che nello stabilire le soglie per il singolo criterio è stata divisa in maniera disomogenea la scala, isolando l'eccellenza, che corrispondeva solamente al valore massimo, e ampliando la classe teorica dei prodotti accettabili. Decisione legittima, stante la volontà, da parte

del Miur e dell'agenzia, di mettere in evidenza i contributi eccellenti. Purtuttavia, non è possibile tralasciare le possibili distorsioni che può aver provocato la sovrastima della votazione "accettabile", anche considerando la successiva fase di attribuzione di una classe di merito sulla base dei punteggi ottenuti sui tre criteri.

Tab. 1 - Corrispondenza tra punteggi espressi da un singolo revisore e classe di merito finale

Classe di merito	*Punteggio sul singolo criterio*	*Punteggio sui tre criteri*
	Soglie	*Soglie*
Eccellente	10	27-30
Elevato	8-9	22-26
Discreto	6-7	16-21
Accettabile	3-5	8-15
Limitato	1-2	3-7

Fonte: Anvur, 2017c, *Appendice B. Linee guida per i revisori*, p. 4

Confrontando le combinazioni teoriche possibili di punteggi sui tre criteri e le soglie prestabilite emerge, infatti, una distribuzione delle classi di merito innegabilmente distorta (Tab. 2). Su 1000 combinazioni, il 41% è riconducibile a un prodotto Discreto, il 39% a un prodotto Accettabile, seguito dai prodotti Elevati (14,5%), Limitati (3,5%) e infine dagli Eccellenti (2%). In altre parole, se si fosse lasciata al caso l'assegnazione dei voti, un prodotto aveva maggiori probabilità di ricadere in alcune classi, rispetto ad altre.

Tab. 2– Distribuzione delle combinazioni di punteggi che restituiscono una determinata classe di merito

Classi di merito	*Combinazioni di punteggi che restituiscono una specifica classe di merito*	*% combinazioni di punteggi che restituiscono una specifica classe di merito*
Limitato	35	3,5%
Accettabile	390	39%
Discreto	410	41%
Elevato	145	14,5%
Eccellente	20	2%
Totale combinazioni	1.000	100%

Questo mancato equilibrio potrebbe aver minato, in alcuni casi, la corrispondenza tra il parere generale del singolo revisore e la definizione della classe di merito corrispondente ai punteggi assegnati. Considerando anche che la scheda di valutazione era rappresentata da una

schermata unica e, di conseguenza, che il revisore poteva vedere immediatamente la classe di merito finale corrispondente al punteggio assegnato, unitamente a una distribuzione disomogenea delle combinazioni di punteggi, il pericolo di un giudizio falsato e distorto è stato particolarmente alto.

A tal proposito, come si è già accennato in precedenza, la condivisione del significato intrinseco di una determinata definizione di qualità della ricerca passa necessariamente per una fase di negoziazione. Come spesso accade anche nel momento in cui viene auto-somministrato un questionario, non essendo presente alcun intervistatore, la specificazione del significato di alcuni termini è un compito che grava interamente sulla formulazione stessa della domanda e, in particolar modo, sulle rispettive alternative di risposta.

Nel momento in cui al revisore veniva mostrata una determinata classe di merito, una volta assegnati i punteggi ad ogni criterio secondo il proprio metro di giudizio, è probabile che il parere personale maturato sul prodotto sottoposto a valutazione non coincidesse con il risultato ottenuto. In casi simili, sono ipotizzabili due reazioni da parte del revisore: rifiutare completamente la classe di merito proposta e, attraverso la manipolazione dei punteggi sui singoli criteri, cambiare il valore finale, oppure, adeguare i propri standard a quelli proposti dall'Anvur e chiudere la scheda. Nel secondo caso, è ipotizzabile che possa esser scattato un meccanismo mentale per il quale il revisore è stato indotto a modificare il proprio metro di giudizio e a convincersi che quel contributo rappresentasse effettivamente quel livello di qualità della ricerca, diverso da quanto egli avrebbe ritenuto originariamente.

Nel momento in cui, tuttavia, la maggior parte delle combinazioni possibili con i punteggi presenti nella scheda di valutazione era riconducibile a classi di merito medio-basse ("discreto" e "accettabile"), c'è stato il rischio che una parte dei revisori fosse stata indotta involontariamente a giudicare relativamente scarsi un gran numero di prodotti, che in origine riteneva di livello superiore. In altre parole, in base alle soglie stabilite per l'assegnazione della classe di merito finale, con molta probabilità, le valutazioni che al revisore potevano apparire positive sono confluite nella parte medio-bassa della scala.

Al fine di approfondire la questione, si è proceduto a una serie di simulazioni, così da ricercare tutte le situazioni d'incongruenza che possono essersi verificate nella valutazione *peer*. In estrema sintesi, ra-

Tab. 3 - Distribuzione teorica sui punteggi minimi (1 revisore)

Impatto attestato/potenziale		*Rigore metodologico*	*Originalità*				
			Limitato	*Accettabile*	*Discreto*	*Elevato*	*Eccellente*
Limitato		*Limitato*	L	L	A	A	A
		Accettabile	L	L	A	A	A
		Discreto	A	A	A	A	D
		Elevato	A	A	A	D	D
		Eccellente	A	A	D	D	D
Accettabile		*Limitato*	L	L	A	A	A
		Accettabile	L	A	A	A	D
		Discreto	A	A	A	D	D
		Elevato	A	A	D	D	D
		Eccellente	A	D	D	D	EL
Discreto		*Limitato*	A	A	A	A	D
		Accettabile	A	A	A	D	D
		Discreto	A	A	D	D	EL
		Elevato	A	D	D	EL	EL
		Eccellente	D	D	EL	EL	EL
Elevato		*Limitato*	A	A	A	D	D
		Accettabile	A	A	D	D	D
		Discreto	A	D	D	EL	EL
		Elevato	D	D	EL	EL	EL
		Eccellente	D	D	EL	EL	EC
Eccellente		*Limitato*	A	A	D	D	D
		Accettabile	A	D	D	D	EL
		Discreto	D	D	EL	EL	EL
		Elevato	D	D	EL	EL	EC
		Eccellente	D	EL	EL	EC	EC

gionando per via tipologica, sono stati costruiti due spazi di attributi, risultanti dall'incrocio dei tre criteri riclassificati secondo i cinque livelli di qualità, ottenendo 125 possibili combinazioni per uno. Nel primo spazio (Tab. 3), per ogni incrocio sono stati riportati i punteggi ottenuti sommando i valori minimi necessari affinché il prodotto rientrasse in quella determinata classe di merito sui singoli criteri (cfr. Tab. 1: 1 per la classe "limitato", 3 per la classe "accettabile", 6 per la classe "discreto", 8 per la classe "elevato" e 10 per la classe "eccellente"); nel secondo spazio (Tab. 4), invece, sono stati riportati gli analoghi punteggi, ottenuti però sommando i valori massimi (cfr. Tab. 1: 2 per la classe "limitato", 5 per la classe "accettabile", 7 per la classe "discreto", 9 per la classe "elevato" e, essendo l'unico punteggio associatogli, 10 per la classe "eccellente"). Sono stati successivamente confrontati i due spazi così ottenuti, al fine di verificarne la similarità o gli eventuali casi incongruenti (Tab. 5).

Tab. 4 - Distribuzione teorica sui punteggi massimi (1 revisore)

Impatto attestato/potenziale		*Rigore metodologico*	*Originalità*				
			Limitato	*Accettabile*	*Discreto*	*Elevato*	*Eccellente*
Limitato		*Limitato*	L	A	A	A	A
		Accettabile	A	A	A	D	D
		Discreto	A	A	D	D	D
		Elevato	A	D	D	D	D
		Eccellente	A	D	D	D	EL
Accettabile		*Limitato*	A	A	A	D	D
		Accettabile	A	A	D	D	D
		Discreto	A	D	D	D	EL
		Elevato	D	D	D	EL	EL
		Eccellente	D	D	EL	EL	EL
Discreto		*Limitato*	A	A	D	D	D
		Accettabile	A	D	D	D	EL
		Discreto	D	D	D	EL	EL
		Elevato	D	D	EL	EL	EL
		Eccellente	D	EL	EL	EL	EC
Elevato		*Limitato*	A	D	D	D	D
		Accettabile	D	D	D	EL	EL
		Discreto	D	D	EL	EL	EL
		Elevato	D	EL	EL	EC	EC
		Eccellente	D	EL	EL	EC	EC
Eccellente		*Limitato*	D	D	D	D	EL
		Accettabile	D	D	EL	EL	EL
		Discreto	D	EL	EL	EL	EC
		Elevato	D	EL	EL	EC	EC
		Eccellente	EL	EL	EC	EC	EC

 ISBN 9788835103219

Tab. 5 - Distribuzione dei casi di congruenza e d'incongruenza, risultata dal confronto tra la classe di merito dei singoli criteri e la classe di merito finale (1 revisore)

			Originalità				
		Rigore metodologico	*Limitato*	*Accettabile*	*Discreto*	*Elevato*	*Eccellente*
Impatto attestato/potenziale	*Limitato*	*Limitato*	C	I	C	C	C
		Accettabile	I	I	C	I	I
		Discreto	C	C	I	I	C
		Elevato	C	I	I	C	C
		Eccellente	C	I	C	C	I
	Accettabile	*Limitato*	I	I	C	I	I
		Accettabile	I	C	I	I	C
		Discreto	C	I	I	C	I
		Elevato	I	I	C	I	I
		Eccellente	I	C	I	I	C
	Discreto	*Limitato*	C	C	I	I	C
		Accettabile	C	I	I	C	I
		Discreto	I	I	C	I	C
		Elevato	I	C	I	C	C
		Eccellente	C	I	C	C	I
	Elevato	*Limitato*	C	I	I	C	C
		Accettabile	C	I	C	I	I
		Discreto	I	C	I	C	C
		Elevato	C	I	C	I	I
		Eccellente	C	I	C	I	C
	Eccellente	*Limitato*	I	I	C	C	I
		Accettabile	I	C	I	I	C
		Discreto	C	I	C	C	I
		Elevato	C	I	C	I	C
		Eccellente	I	C	I	C	C

I risultati di questa simulazione, riportati in Tab. 5, sono abbastanza singolari. Applicando le soglie prestabilite, 65 su 125 combinazioni possibili riproducono incongruenze. Si tratterebbe di prodotti che, pur rientrando in uno spazio di attributi e nell'altro nelle stesse classi di merito per quanto riguarda i punteggi sui singoli criteri, ottengono un

risultato finale diverso di almeno una classe, a seconda che la somma dei punteggi sia data dai valori minimi o massimi stabiliti dalle soglie. Inoltre, in alcuni casi, le combinazioni non riproducono solo incongruenze, ma anche paradossi. Per esempio, per rientrare nella classe di merito Eccellente (soglie 27-30), un prodotto può essere giudicato Elevato su tutti e tre i criteri, ottenendo un punteggio pari a 9 sulle tre domande (9+9+9=27); ma se fosse giudicato Elevato, con votazione pari a 8, su due dimensioni ed Eccellente su una sola, paradossalmente rientrerebbe nella classe di merito inferiore (8+8+10=26).

Queste poche considerazioni sembrano bastare per legittimare una modifica delle procedure di sintesi del giudizio. Le soluzioni auspicabili per risolvere l'incongruenza tra la classe di merito assegnata in base ai criteri e quella finale sono principalmente due: una volta classificati i prodotti sulla base dei singoli criteri, si potrebbe procedere tramite la costruzione di uno spazio di attributi per l'attribuzione della classe di merito finale; oppure si potrebbe pensare di tralasciare la classificazione sul singolo criterio e, di conseguenza, non stabilire alcuna soglia, procedendo direttamente alla somma dei punteggi e alla successiva classificazione sulla base delle soglie finali.

La prima strada non richiede alcuna operazione matematica e renderebbe la procedura apparentemente immune dai paradossi appena descritti. Anche da un punto di vista metodologico si intravedono alcuni vantaggi: stante il rapporto che i criteri hanno tra di loro nella copertura semantica del concetto di qualità della ricerca, si può pensare infatti a una soluzione che tenga conto non solo del loro sommarsi, ma anche delle combinazioni possibili (Agnoli, 1994, p. 167; Faggiano, 2012, p. 83). Inoltre, il fatto che la definizione operativa del concetto originario abbia portato alla costruzione di variabili quasi-cardinali non impedisce in alcun modo di trattarle come variabili categoriali ed effettuare operazioni di sintesi per queste più congeniali.

In letteratura, uno spazio di attributi viene definito come «un insieme di oggetti, caratterizzato mediante una serie di dati, che saranno definiti in termini di punti all'interno di uno spazio; quest'ultimo dovrà avere tante dimensioni quanti sono i dati necessari per caratterizzare ciascuno degli oggetti in questione» (Lazarsfeld, 1966, tr. it., 2001, p. 154).

In questo modo, tuttavia, risulterebbero ben 125 tipi possibili, ren-

 ISBN 9788835103219

dendo necessaria una riduzione dello spazio di attributi, ossia l'aggregazione di alcuni tipi, secondo regole logico-procedurali abbastanza precise. Le difficoltà che si incontrano in questa procedura non sono certo trascurabili, prima fra tutte l'inevitabile perdita d'informazioni e la natura negoziale della stessa, ciononostante rappresenterebbe uno stimolo necessario all'approfondimento e al miglioramento del processo di sintesi dei giudizi valutativi.

Diverse sono le strade possibili da percorrere quando un ricercatore si trova a dover ridurre uno spazio di attributi. Generalmente, i criteri principali da rispettare sono la vicinanza semantica dei tipi che dovranno formare una categoria più generale e il numero di casi (Marradi, 2007, p. 185). Tuttavia, mentre nel caso di una ricerca sociale, volta eventualmente a corroborare teorie, controllare ipotesi o esplorare fenomeni, assumere il numero di casi che rientrano in una determinata categoria come criterio per la riduzione dei tipi è un procedimento quasi necessario, altrimenti risulterebbe pressoché impossibile lavorare in sede di analisi, nel caso della riduzione delle classi di merito per la Vqr, una tale condizione non sembra assumere alcun valore. È ragionevole lavorare, perciò, seguendo una logica principalmente semantica e tentare di aggregare quei tipi dotati di una maggiore prossimità.

Osservando la Tab. 6, che ripropone le diverse procedure di riduzione suggerite da Lazarsfeld e Barton (1951; tr. it., 1961, pp. 265-272), la strada più idonea sembrerebbe essere la riduzione *pragmatica*, la quale può essere condotta anche prima della raccolta dei dati ed è guidata dagli scopi della ricerca (Fideli, 2001, pp. 126-127).

Tab. 6 - Procedure di riduzione di uno spazio di attributi e relative definizioni

Procedure di riduzione	Definizioni
Semplificazione delle dimensioni	Consiste nel ridurre variabili continue in classi ordinate o un gruppo di classi in una dicotomica
Pragmatica	Certe combinazioni vengono concentrate in una classe in vista degli scopi della ricerca
Funzionale	Certe combinazioni possono essere completamente eliminate, oppure verificarsi così raramente che non occorre fissare una classe speciale per queste
Costruzione di un indice numerico (arbitraria numerica)	Consiste nell'attribuire ad ogni categoria un certo valore per ogni dimensione e quindi nel sommare i valori così ottenuti

Fonte: Fideli, *La costruzione di un indice tipologico: criteri semantici, numerici ed empirici*, p. 126

La procedura risulterebbe semplificata nel momento in cui fossero aggregati tutti quei tipi che derivano dall'incrocio delle stesse classi di merito, indipendentemente dal criterio che vi rientra; infatti, i tipi che ne derivano passerebbero dall'essere 125 a 35. Ciò avviene perché, data l'equità tra le dimensioni nel contribuire alla specificazione del significato di qualità della ricerca, l'attenzione può focalizzarsi sulle diverse combinazioni di classi di merito, rendendo irrilevante quale criterio rientri in quale classe. Quindi, per esempio, la combinazione "Limitato-Accettabile-Elevato" si ripeterebbe ben sei volte all'interno dello spazio di attributi costruito sui tre criteri, ma date le condizioni appena descritte, basterebbe prendere in considerazione il tipo che ne deriva una sola volta; analogamente, è possibile ridurre le altre combinazioni che si ripropongono dalle 3 alle 6 volte, a eccezione dei casi in cui vi è completa convergenza tra i criteri.

Stanti le considerazioni svolte fino ad ora, tuttavia, sorgono alcuni problemi di non poca importanza relativi all'utilizzo di tale procedura. Innanzitutto, anche riducendo drasticamente i tipi, il numero che ne deriva (35) risulterebbe ancora troppo elevato. Maneggiare e definire un numero così alto d'ipotetiche categorie rischia di divenire un compito troppo gravoso per un'eventuale rivisitazione delle procedure, in vista del prossimo esercizio di valutazione.

In secondo luogo, la riflessione fin qui condotta si è basata sulla scheda effettivamente utilizzata per la Vqr 2011-2014, la quale, tuttavia, come si è già visto nel capitolo precedente, risulta avere non pochi problemi. La definizione operativa dei criteri del concetto di qualità effettuata manca, infatti, di alcuni passaggi necessari all'individuazione di referenti empirici più specifici e funzionali alla rilevazione del concetto più generale. Senza entrare nuovamente nel merito della questione[3], basti ricordare che solamente la prima domanda della scheda di valutazione, così come è stata formulata, è scomponibile in sei sotto-domande, ognuna riferita ad un aspetto diverso dell'originalità; analogamente, il rigore metodologico e l'impatto attestato/potenziale possono essere rilevati, rispettivamente, attraverso un numero più ampio di *items*.

A questo punto, sembrano evidenti le grandi difficoltà (se non l'impossibilità) di applicazione di una procedura come quella della riduzione dello spazio di attributi, il quale si estenderebbe a tal punto da

[3] Per una discussione più approfondita della questione si vedano i Capitoli 2 e 4.

divenire troppo complicato da gestire. Al fine di evitare le incongruenze riscontrate, è auspicabile procedere per una strada diversa, sulla scia della prima soluzione proposta nelle pagine precedenti: tralasciare la classificazione sulla base di soglie prestabilite dei singoli criteri e procedere direttamente alla costruzione dell'indice, tramite sommatoria dei punteggi, e definire la classe di merito in base al solo risultato finale.

Riprendendo la proposta di scheda di valutazione effettuata nel capitolo precedente (cfr. capitolo 4), i punteggi che un revisore può fornire sui singoli criteri avrebbero, tuttavia, intervalli differenti: l'originalità ricadrebbe in un intervallo che va da un minimo di 6 ad un massimo di 60, il rigore metodologico varierebbe tra un minimo di 5 ed un massimo di 50, mentre l'impatto attestato/potenziale tra un minimo di 3 ed un massimo di 30. Prima di procedere alla sommatoria finale, è pertanto necessario effettuare una semplice media aritmetica dei punteggi assegnati sui singoli criteri, in modo tale da evitare che criteri con *range* più ampi possano pesare maggiormente sul risultato finale. In questo modo, i punteggi finali che si possono ottenere sui singoli criteri tornerebbero ad avere lo stesso identico intervallo, ossia 1-10. Procedendo a questo punto con la semplice sommatoria si ricadrebbe tuttavia nelle stesse problematiche affrontate nelle pagine precedenti, a meno che non vengano apportate modifiche alle soglie che stabiliscono le classi di merito.

Nella Tab. 7 sono riportate sia le soglie utilizzate per la Vqr 2011-2014, sia le possibili ulteriori soglie per riequilibrare i punteggi corrispondenti alle classi di merito.

La soluzione prevede una semplice ridistribuzione bilanciata dei valori, tale per cui le probabilità per un prodotto di ricadere in una determinata classe di merito siano distribuite in maniera normale. In altre parole, la percentuale di combinazioni che restituiscono una classe di merito va a diminuire agli estremi e ad aumentare verso le classi centrali, indipendentemente dalla "direzione". Andando più nello specifico, sono state ritoccate le soglie di quattro classi, diminuendo il *range* da 8 punteggi a 6, per la classe "Accettabile", e aumentando di un punto il *range* delle classi "Eccellente", "Elevato" e "Discreto"; in questo modo, le classi estreme occupano entrambe cinque punti della scala, mentre le altre tre ne occupano sei.

Tab. 7 - Corrispondenza tra punteggi e classi di merito e distribuzione delle combinazioni di punteggi che restituiscono una determinata classe di merito

Classe di merito	*Vqr 2011-2014*			*Possibili ulteriori soglie*		
	Soglie punteggi dei tre criteri (1 revisore)	*Combinazioni di punteggi che restituiscono una specifica classe di merito*	*% combinazioni di punteggi che restituiscono una specifica classe di merito*	*Soglie punteggi dei tre criteri (1 revisore)*	*Combinazioni di punteggi che restituiscono una specifica classe di merito*	*% combinazioni di punteggi che restituiscono una specifica classe di merito*
Eccellente	27-30	20	2%	26-30	35	3,5%
Elevato	22-26	145	14,5%	20-25	248	24,8%
Discreto	16-21	410	41%	14-19	434	43,4%
Accettabile	8-15	390	39%	8-13	248	24,8%
Limitato	3-7	35	3,5%	3-7	35	3,5%

Fonte: Anvur, 2017c, *Appendice B. Linee guida per i revisori*, p. 4 e rielaborazione propria

Per quanto riguarda, invece, le incongruenze tra le classi di merito in cui possono ricadere i singoli criteri e la classe finale (Tab. 3, 4 e 5), in questo modo tale problema non si porrebbe, in quanto verrebbe direttamente tralasciato il passaggio.

Non si vuole certo suggerire di nascondere semplicemente la questione, bensì far ragionare sulla poca, o quasi nulla, funzionalità di tale fase all'interno dell'intero processo. In nessun documento ufficiale vi è traccia dell'utilità di una classificazione dei prodotti sui singoli criteri, ai fini dell'assegnazione della classe di merito finale; inoltre, la scheda di valutazione, così come era impostata, non sembra neanche aver salvato traccia di tale passaggio, in quanto il sistema restituiva e conservava solamente la classe di merito finale. L'unico beneficio che si intravede si trova nel poter semplificare al revisore il lavoro cognitivo di dover associare il proprio parere ad un numero della scala proposta, potendo confrontare le soglie prestabilite con le corrispettive classi di merito. Tuttavia, nei documenti consultabili dai revisori prima della chiusura dei lavori non vi è traccia di alcun riferimento alla corrispondenza tra classi e punteggi. L'unico accenno alla classificazione dei singoli criteri è in appendice ai rapporti di area finali, pubblicati solamente nel mese di febbraio 2017.

5.3. L'attribuzione della classe di merito finale

I giudizi espressi dai due revisori, con tutte le contraddizioni viste fino ad ora, una volta inviati al Gev, sono stati "trasformati" in una delle cinque classi di merito (Anvur, 2017c, p. 5). Questo è il passaggio più delicato e importante dell'intero esercizio valutativo, poiché ad ogni classe è stato successivamente associato un peso che contribuiva al calcolo degli indicatori, utilizzati per stilare la graduatoria dei Dipartimenti universitari (Anvur, 2015, pp. 21-22; Anvur, 2017, p. 32). Pur essendo un passaggio cruciale, nei documenti stilati dai Gev tuttavia non è presente alcun tipo d'informazione sulle procedure e sui criteri utilizzati per sintetizzare i giudizi dei revisori.

Plausibilmente, sono stati confrontati i punteggi ottenuti da ciascun prodotto e, una volta valutato il loro grado di convergenza, è stato attribuito ad una classe di merito ovvero sottoposto ad altra revisione o al gruppo di consenso. Osservando la Tab. 8, che riporta tutte le possibili situazioni in cui i Gev si possono essere ritrovati, è possibile notare che in 10 casi, al di là dei giudizi perfettamente concordanti (rappresentati dalla diagonale principale), gli Esperti Valutatori hanno dovuto prendere delle decisioni su come procedere. Decisioni che avrebbero dovuto essere argomentate approfonditamente in ragione dell'ampiezza della divergenza.

Tab. 8 - Sintesi dei giudizi di due revisori (simulazione)

		Punteggio assegnato dal revisore A				
		3-7	*8-15*	*16-21*	*22-26*	*27-30*
Punteggio assegnato dal revisore B	*3-7*	L	N	N N	N N N	N N N N
	8-15	N	A	N	N N	N N N
	16-21	N N	N	D	N	N N
	22-26	N N N	N N	N	EL	N
	27-30	N N N N	N N N	N N	N	EC

A questo punto, sorgono spontanee alcune domande: quale ragionamento c'è stato dietro alla decisione finale di attribuzione di una classe di merito ai prodotti? Quale strada è stata intrapresa dai Gev? È possibile solamente fare delle ipotesi.

Anche in questo caso, le opzioni possono esser essenzialmente due: procedere alla definizione dei tipi che risultano dall'incrocio delle due revisioni ovvero sommare i punteggi finali e stabilire delle soglie per la corrispondenza con le classi di merito. Per quanto riguarda la prima soluzione, la procedura sarebbe estremamente semplificata rispetto a quella proposta precedentemente per la sintesi del giudizio di un solo revisore, poiché, assumendo l'equidistanza tra le classi, si tratterebbe di definire solamente cinque tipi: i casi di giudizi concordanti e i casi di giudizi discordanti di una, due, tre e quattro classi. In altre parole, sarebbe necessario solamente stabilire le regole per l'assegnazione della classe nel caso di controversie. La seconda strada percorribile, invece, prevederebbe la costruzione di un indice additivo, auspicabilmente normalizzato. Riprendendo, quindi, la proposta effettuata poc'anzi, si potrebbe pensare di eseguire una semplice media aritmetica[4], sommando i punteggi e dividendo successivamente per due. In questo modo, l'intervallo risulterebbe essere sempre ricompreso tra 3 e 30, così da permettere l'utilizzo delle stesse soglie applicate al giudizio del singolo revisore (Par. 2).

C'è da chiedersi, tuttavia, in questo secondo caso, quale sarebbe l'utilità dell'attribuzione di una classe di merito da parte dei singoli revisori, se infine fosse possibile utilizzare esclusivamente il punteggio. Il nodo della questione, secondo chi scrive, si trova nella necessità di ancorare semanticamente una scala di punteggi standardizzata utilizzata da revisori diversi, ognuno potenzialmente con un metro di giudizio personale. Per agevolare la convergenza tra valutazioni dei revisori e i livelli di qualità concettualizzati dall'agenzia risulterebbe pertanto funzionale la condivisione del significato del punteggio finale che viene attribuito ai prodotti. In altre parole, rendere esplicita la classe di merito corrispondente al giudizio al singolo revisore avrebbe il vantaggio, da un punto di vista pratico, di facilitare la comprensione del significato attribuito ai valori numerici, seppur potrebbe risultare un passaggio privo di utilità da un punto di vista tecnico e procedurale.

A favore dell'opzione di costruzione di un indice additivo per sintetizzare le valutazioni vi è anche l'applicazione di una procedura quasi identica nella Vqr 2004-2010; in quel caso, l'agenzia aveva predisposto delle soglie anche per la sintesi dei giudizi dei due revisori,

[4] Come verrà esposto nelle pagine che seguono, l'espediente della media aritmetica è stato utilizzato dal Gev10 per risolvere i casi di giudizi discordanti tra i revisori (Anvur, 2017h, p. 10).

 ISBN 9788835103219

prevedendo casi anche di tre o più revisioni. Non è chiaro, pertanto, cosa abbia portato a ignorare questo passaggio nella Vqr 2011-2014.

In merito alla gestione delle controversie, come già argomentato nel capitolo 3, ci si imbatte in un grande problema di *accountability*: nel d.m. n. 458 del 2015 era espressamente specificato che i Gev avevano il compito di redigere il rapporto finale, illustrando, tra le altre cose, la procedura adottata per la risoluzione di eventuali conflitti di valutazione da parte dei revisori. Tuttavia, nel rapporto finale del Gev 14 i riferimenti alle controversie sono pressoché nulli. Viene sommariamente e in più punti assunto l'utilizzo sia di terze revisioni (n. 130) che dei *consensus group* (n. 143), senza alcuna indicazione sulla composizione di quest'ultimi, sul grado di discordanza che le revisioni dovevano presentare affinché si ricorresse alla terza revisione e sulle modalità di gestione e risoluzione delle controversie da parte degli Ev stessi.

Tab. 9 - Numero e percentuali di revisioni peer concordanti e discordanti per 1, 2, 3 e 4 classi per area

Area	*Prodotti*	*Conc.*	*%*	*1 classe*	*%*	*2 classi*	*%*	*3 classi*	*%*	*4 classi*	*%*
1	2.356	1.114	47,28	927	39,35	262	11,12	47	1,99	6	0,25
2	1.291	458	35,48	633	49,03	166	12,86	29	2,25	5	0,39
3	1.394	543	38,95	660	47,35	165	11,84	25	1,79	1	0,07
4	1.257	478	38,03	531	42,24	215	17,10	29	2,31	4	0,32
5	2.183	825	37,79	981	44,94	303	13,88	68	3,11	6	0,27
6	3.731	1.212	32,48	1.675	44,89	679	18,20	150	4,02	15	0,40
7	2.463	952	38,65	1.049	42,59	380	15,43	72	2,92	10	0,41
8a	3.433	1.208	35,19	1.519	44,25	566	16,49	119	3,47	21	0,61
8b	996	430	43,17	387	38,86	151	15,16	24	2,41	4	0,40
9	3.346	1.129	33,74	1.557	46,53	534	15,96	116	3,47	10	0,30
10	8.720	3.024	34,68	3.920	44,95	1.379	15,81	339	3,89	58	0,67
11a	5.956	2.213	37,16	2.653	44,54	873	14,66	192	3,22	25	0,42
11b	868	286	32,95	392	45,16	148	17,05	38	4,38	4	0,46
12	8.431	2.934	34,80	3.857	45,75	1.333	15,81	276	3,27	31	0,37
13	2.662	909	34,15	1.105	41,51	502	18,86	139	5,22	7	0,26
14	2.953	995	33,69	1.254	42,47	532	18,02	156	5,28	16	0,54
Totale	*52.040*	*18.710*	*35,95*	*23.100*	*44,39*	*8.188*	*15,73*	*1.819*	*3,50*	*223*	*0,43*

Fonte: Anvur, 2017, *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale Anvur. Parte Prima: Statistiche e risultati di compendio*, p. 27

Tale mancanza non è certamente da sottovalutare, considerando l'elevato numero di prodotti che hanno ottenuto valutazioni discordanti, che nella sola Area 14 rappresenta il 64% dei prodotti sottomessi a *peer review* (Tab. 9).

In particolare, lascia perplessi la grande differenza tra il numero di prodotti sottoposti a terza revisione o al *consensus group* e il numero di controversie. Considerando anche i casi discordanti di una sola classe, la differenza nell'Area 14 è di 1.685 casi per cui non è chiara la procedura di risoluzione della controversia, ma anche considerando solo i casi da due classi in poi, la differenza non è trascurabile (n. 431).

La scarsa trasparenza per questa particolare fase del processo di valutazione *peer* non è solo un problema dell'Area 14. Leggendo i rapporti di area degli altri gruppi di valutazione sono rintracciabili informazioni in merito solamente per tre Gev.

Nel rapporto del Gev 06, per esempio, vengono specificate tutte le informazioni che sono state prese in considerazione dagli Ev nel caso in cui siano stati chiamati a ricondurre in una determinata classe di merito un prodotto che ha ricevuto giudizi discordi dai due revisori. Nel caso di articoli su rivista, quindi, hanno considerato anche la classificazione su base bibliometrica, nel caso di pubblicazioni prive di tali indicatori, invece, hanno tenuto conto delle caratteristiche dell'edizione (collana editoriale, il comitato editoriale, procedure di revisione, diffusione e prestigio delle pubblicazioni dell'editore, recensioni ricevute dalla pubblicazione, ecc.) (Anvur, 2017f, p. 18). I Gev 07 e 10, invece, hanno lavorato distinguendo diversi gradi di controversie. Nel caso in cui le valutazioni fossero state diverse di una sola classe, avrebbero deciso la classe di merito finale i due membri che avevano gestito il prodotto (Anvur, 2017g, p. 10), mentre il Gev 10 ha optato per il calcolo della media dei punteggi assegnati dai due revisori (Anvur, 2017h, p. 10). Nel caso in cui, invece, le valutazioni fossero state distanti di due o più classi, entrambi i Gev hanno optato per la costituzione di un gruppo di consenso al loro interno che avrebbe proposto un giudizio finale, tramite la "metodologia del *consensus group*"; il Gev 07, inoltre, riporta ulteriori informazioni sulla formazione del gruppo stesso, composto da 5 membri, tra i quali gli stessi che hanno gestito il prodotto.

La presenza di un numero così elevato di prodotti "controversi" può essere vista come il sintomo di una qualche mancanza nella progetta-

zione dell'intero processo di valutazione, o quantomeno di un problema cui involontariamente non è stata dedicata l'attenzione opportuna. La questione diventa di rilevante importanza nel momento in cui, allargando l'orizzonte anche alle altre aree disciplinari, comprese le bibliometriche, l'ammontare dei casi di controversie non sembra variare significativamente. Il fenomeno dei giudizi discordanti e la mancata convergenza delle valutazioni *peer* non sembrano pertanto essere stati una prerogativa delle aree non bibliometriche, né tantomeno della sola area delle scienze politiche e sociali, spesso e da molti considerata un'area poco pacifica, vista la presenza al suo interno di diverse scuole di pensiero e approcci alla disciplina.

Tab. 10 - Scarto percentuale dei prodotti valutati tramite peer review *che hanno ottenuto due valutazioni discordani di almeno una classe di merito*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8a*	*8b*	*9*	*10*	*11a*	*11b*	*12*	*13*	*14*
1	0	-11,8	-8,3	-9,3	-9,5	-14,8	-8,6	-12,1	-4,1	-13,5	-12,6	-10,1	-14,3	-12,5	-13,1	-13,6
2	11,8	0	3,5	2,6	2,3	-3	3,2	-0,3	7,7	-1,7	-0,8	1,7	-2,5	-0,7	-1,3	-1,8
3	8,3	-3,5	0	-0,9	-1,2	-6,5	-0,3	-3,8	4,2	-5,2	-4,3	-1,8	-6	-4,2	-4,8	-5,3
4	9,3	-2,6	0,9	0	-0,2	-5,5	0,6	-2,8	5,1	-4,3	-3,3	-0,9	-5,1	-3,2	-3,9	-4,3
5	9,5	-2,3	1,2	0,2	0	-5,3	0,9	-2,6	5,4	-4,1	-3,1	-0,6	-4,8	-3	-3,6	-4,1
6	14,8	3	6,5	5,5	5,3	0	6,2	2,7	10,7	1,3	2,2	4,7	0,5	2,3	1,7	1,2
7	8,6	-3,2	0,3	-0,6	-0,9	-6,2	0	-3,5	4,5	-4,9	-4	-1,5	-5,7	-3,9	-4,5	-5
8a	12,1	0,3	3,8	2,8	2,6	-2,7	3,5	0	8	-1,4	-0,5	2	-2,2	-0,4	-1	-1,5
8b	4,1	-7,7	-4,2	-5,1	-5,4	-10,7	-4,5	-8	0	-9,4	-8,5	-6	-10,2	-8,4	-9	-9,5
9	13,5	1,7	5,2	4,3	4,1	-1,3	4,9	1,4	9,4	0	0,9	3,4	-0,8	1,1	0,4	0
10	12,6	0,8	4,3	3,3	3,1	-2,2	4	0,5	8,5	-0,9	0	2,5	-1,7	0,1	-0,5	-1
11a	10,1	-1,7	1,8	0,9	0,6	-4,7	1,5	-2	6	-3,4	-2,5	0	-4,2	-2,4	-3	-3,5
11b	14,3	2,5	6	5,1	4,8	-0,5	5,7	2,2	10,2	0,8	1,7	4,2	0	1,9	1,2	0,7
12	12,5	0,7	4,2	3,2	3	-2,3	3,9	0,4	8,4	-1,1	-0,1	2,4	-1,9	0	-0,7	-1,1
13	13,1	1,3	4,8	3,9	3,6	-1,7	4,5	1	9	-0,4	0,5	3	-1,2	0,7	0	-0,5
14	13,6	1,8	5,3	4,3	4,1	-1,2	5	1,5	9,5	0	1	3,5	-0,7	1,1	0,5	0

Fonte: Rielaborazione dei dati a partire da Anvur, 2017, *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale Anvur. Parte Prima: Statistiche e risultati di compendio*, p. 27

Nella Tab. 10 sono riportati gli scarti percentuali tra tutte le aree disciplinari delle valutazioni *peer* discordanti di almeno una classe di

merito; a partire dal numero di prodotti che hanno ottenuto giudizi diversi da parte dei due revisori, sono state quindi calcolate le distanze, in termini percentuali, per ogni area dalle altre, per verificare se vi siano state differenze significative.

La tabella è stata costruita in modo da esser letta più agevolmente per riga, nella quale sono riportati gli scarti, negativi e positivi, dell'area posta in riga rispetto all'area posta in colonna, il tutto in maniera intuitiva. Nel caso in cui il valore risulti positivo, nell'area in riga vi è stato un numero di valutazioni discordanti maggiore rispetto all'area posta in colonna. Viceversa, se il valore riportato è negativo, vuol dire che l'area ha avuto un numero medio di casi controversi inferiore rispetto all'area posta a confronto. Essendo una tabella simmetrica, nella diagonale principale, che rappresenta l'incrocio di un'area con se stessa, sono riportati sempre valori pari a zero.

Osservando le distanze tra le aree, le uniche che si distinguono, presentando, sempre o quasi, scarti negativi superiori al 5% rispetto alle altre, sono l'Area 1 (Scienze matematiche e informatiche) e l'Area 8b (Ingegneria civile). In tutti gli altri casi, i valori che rappresentano la distanza in termini di percentuale di valutazioni discordanti non sembrano raggiungere livelli che lascino intendere che il fenomeno abbia interessato una disciplina in maniera maggiore rispetto a un'altra. L'Area 14, in particolare, sembra allinearsi alle altre aree non bibliometriche, in taluni casi con uno scarto prossimo anche allo zero. L'ipotesi secondo cui, quindi, il livello di discordia tra revisori sia stato particolarmente accentuato in determinate aree piuttosto che in altre sembrerebbe disconfermata. Tuttavia, la possibilità che vi siano state differenze ragguardevoli tra aree bibliometriche e non ha necessitato un approfondimento. Per questo motivo, sono stati calcolati anche gli scarti percentuali delle due macro-aree, e corrispondenti sotto aree, rispetto al totale (Tab. 11).

La differenza in termini di casi di valutazioni discordanti tra macro-aree risulta, in questo caso, decisamente ridimensionata. Mentre la macro-area non bibliometrica ha avuto solamente lo 0,78% di valutazioni discordanti in più rispetto al totale dei prodotti sottoposti a *peer review*, la macro-area bibliometrica ha uno scarto in negativo di appena -1,02%. Inoltre, osservando nel dettaglio gli scarti tra il totale e le singole aree, al di là dei due casi eccezionali evidenziati prima, rappresentati dalle aree 1 e 8b, nessun valore raggiunge livelli significativi, né in negativo né in positivo.

Tab. 11– Distribuzione di frequenze delle valutazioni concordanti e discordanti, scarto percentuale dei prodotti valutati tramite peer review *che hanno ottenuto due valutazioni discordanti di almeno una classe di merito rispetto alla macro-area e al totale, per aree bibliometriche e non bibliometriche*

Area disciplinare	*Totale prodotti inviati a peer review*	*Valutazioni concordanti*	*%*	*Valutazioni discordanti di almeno una classe*	*%*	*Scarto percentuale dalla macro-area*	*Scarto percentuale dal totale*
1	2.356	1.114	47,28	1.242	52,72	-10,31	-11,33
2	1.291	458	35,48	833	64,52	1,5	0,48
3	1.394	543	38,95	851	61,05	-1,98	-3
4	1.257	478	38,03	779	61,97	-1,06	-2,07
5	2.183	825	37,79	1.358	62,21	-0,82	-1,84
6	3.731	1.212	32,48	2.519	67,52	4,49	3,47
7	2.463	952	38,65	1.511	61,35	-1,68	-2,7
8b	996	430	43,17	566	56,83	-6,2	-7,22
9	3.346	1.129	33,74	2.217	66,26	3,23	2,21
11b	868	286	32,95	582	67,05	4,02	3
13	2.662	909	34,15	1.753	65,85	2,82	1,81
Totale aree bibliometriche	*22.547*	*8.336*	*36,97*	*14.211*	*63,03*		*-1,02*
8a	3.433	1.208	35,19	2.225	64,81	-0,01	0,77
10	8.720	3.024	34,68	5.696	65,32	0,5	1,27
11a	5.956	2.213	37,16	3.743	62,84	-1,98	-1,2
12	8.431	2.934	34,8	5.497	65,2	0,37	1,15
14	2.953	995	33,69	1.958	66,31	1,48	2,26
Totale aree non bibliometriche	*29.493*	*10.374*	*35,17*	*19.119*	*64,83*		*0,78*
Totale	*52.040*	*18.710*	*35,95*	*33330*	*64,05*		

Fonte: Rielaborazione dei dati a partire da Anvur, 2017, *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale Anvur. Parte Prima: Statistiche e risultati di compendio*, p. 27

A questo punto è chiaro che il fenomeno delle controversie abbia rappresentato un problema trasversale per tutte le aree e che non fosse dovuto a caratteristiche particolari di una disciplina o dei rispettivi membri della comunità scientifica di riferimento. In quasi tutte le aree, oltre il 60% dei prodotti sottoposti a *peer review* ha ricevuto due valutazioni differenti da parte dei revisori e anche nelle aree più "virtuose" tale numero rappresenta oltre la metà dei casi.

Arrivati a questo punto della discussione appare alquanto arduo stabilire quale sia stata la radice del problema. Si è visto, infatti, come a

partire dalla definizione stessa del concetto di qualità tramite le tre dimensioni (capitolo 2) siano state prese decisioni alquanto discutibili che possono aver avuto ripercussioni a valanga su tutto il processo valutativo. I criteri ambigui e poco selettivi di costruzione della lista dei revisori possono aver provocato un'autoselezione degli stessi (capitolo 3). Infine, lo strumento di valutazione *peer* dei prodotti è stato costruito senza tener conto di alcuni accorgimenti metodologici (capitolo 4). Senza ignorare gli sforzi di miglioramento rispetto alla Vqr 2004-2010, sembra esser però mancata una particolare attenzione ad alcuni dettagli, piccoli e grandi, che hanno reso distorta l'intera procedura con evidenti ripercussioni sui risultati.

In particolare, rispetto alla procedura di valutazione *peer* merita di esser rimarcata l'importanza di una problematica già affrontata nei capitoli precedenti. Come avviene, per esempio, nell'analisi del contenuto, la quale anche opera su proprietà qualitative, uno dei momenti più delicati riguarda la fase di raccolta dei dati, a causa principalmente dei possibili errori dovuti alla mancata condivisione di codici interpretativi da parte degli analisti; in altre parole, in taluni casi, ha luogo un problema di condivisione del significato del concetto da rilevare *inter* e *intra* rilevatore: «un analista del contenuto rileva lo stesso materiale nel medesimo modo con due rilevazioni successive? E poi: due diversi analisti rilevano lo stesso materiale nello stesso modo nella stessa unità di tempo?» (Nobile, 1997, pp. 77-78).

Trasponendo questi interrogativi alla Vqr, bisognerebbe quindi chiedersi: un revisore valuta allo stesso modo un prodotto in momenti diversi? Due diversi revisori esprimono un parere sullo stesso prodotto, condividendo lo stesso metro di giudizio?

Per quanto riguarda il primo interrogativo è bene ribadire che un controllo sulla stabilità di giudizio può avvenire solamente in un momento antecedente all'esercizio valutativo, per il semplice motivo che i revisori sono chiamati a esprimere un giudizio sempre su prodotti diversi. Come già suggerito nel capitolo 4, pensare di progettare una fase di addestramento, innanzitutto, all'utilizzo della scheda di valutazione rimane la soluzione ideale per arginare la volubilità delle valutazioni. Un lavoro analogo andrebbe effettuato anche per rendere più esplicito e familiare il concetto stesso di qualità della ricerca a tutti gli attori coinvolti (revisori, esperti valutatori, addetti), lavorando nella

direzione di una definizione delle sue dimensioni più chiara, condivisibile e priva di ambiguità terminologiche, temporali e concettuali (capitolo 2).

Il secondo interrogativo, invece, potrebbe essere affrontato a valle dell'esercizio appena concluso, grazie all'enorme mole d'informazioni che ha generato e congiuntamente al problema delle valutazioni *peer* discordanti. Sarebbe, pertanto, auspicabile da parte dell'agenzia progettare una riflessione approfondita in merito alla questione, seguendo una logica procedurale che si avvicini all'analisi dei casi devianti (capitolo 6).

I risultati di un'analisi simile risulterebbero funzionali al miglioramento dell'intero esercizio valutativo, dalla possibilità di una chiarificazione dell'intensione del concetto di qualità della ricerca, finanche al controllo della validità degli indicatori e dell'affidabilità della definizione operativa (Mauceri, 2008).

5.4. Conclusioni

La fase cruciale di classificazione dei prodotti in base alla loro qualità, come si è visto nel corso del capitolo, è stata caratterizzata da alcune lacune procedurali, con evidenti ripercussioni sulla qualità dei risultati. Seppur ravvisabile un miglioramento nella specificazione di alcuni aspetti delle procedure, il tentativo di spalmare a tutti gli ambiti disciplinari le stesse prassi e regole ha, involontariamente o meno, reso ambigue alcune operazioni procedurali.

Con la poca trasparenza riservata ad alcune fasi del processo di valutazione non è possibile tuttavia (e non si avrebbe neanche la pretesa di farlo) fornire alcuna procedura propositiva di risoluzione del problema che possa funzionare in assoluto. Ciò che si vuole rimarcare è, tuttavia, la *costruttività* delle critiche qui esposte.

Il nodo cruciale da sciogliere, in vista del prossimo esercizio di valutazione della qualità della ricerca, è la definizione del concetto stesso che si vuole rilevare e, conseguentemente, dei diversi livelli in cui può esser rappresentato. In altre parole, la questione da dirimere primariamente è la specificazione del significato, attualmente implicito, delle classi di merito in relazione all'ambito disciplinare specifico. Non che

si voglia ignorare il notevole sforzo effettuato nell'apportare modifiche migliorative alle definizioni lessicali delle classi di merito. Anzi, come si è avuto modo di vedere, nella Vqr 2011-2014 vi è stata una notevole attenzione nella chiarificazione dei gradi di qualità in cui un prodotto poteva ricadere.

Ciò che si vuole rimarcare sono le importanti ripercussioni che una maggior chiarificazione dei livelli di qualità può offrire all'intero esercizio. Dalla stesura stessa della scheda di rilevazione alla formulazione del giudizio finale, riuscire a costruire un senso condiviso delle regole, dei criteri, del significato stesso dell'eccellenza o della scarsità di un prodotto di ricerca può migliorare e chiarificare il senso stesso della valutazione della qualità della ricerca.

 ISBN 9788835103219

6. Fare ricerca sulla Vqr

di Fabrizio Martire

L'obiettivo di questo breve capitolo è proporre idee di ricerca sulla Vqr, relativamente alle finalità per cui è stata introdotta dal Miur (vedi il capitolo 1 di questo volume), agli attori coinvolti nel processo di valutazione (vedi il capitolo 3), agli strumenti di valutazione (capitolo 4), al giudizio sui prodotti (capitolo 5).

Nel primo capitolo Palmieri sottolinea come, a norma di legge, una quota sempre più consistente del Fondo di Finanziamento Ordinario alle università pubbliche (Ffo), la cosiddetta quota premiale, sia redistribuita tra gli atenei e i dipartimenti universitari in considerazione degli esiti della valutazione dei prodotti della ricerca; nel 2017 il Miur ha destinato ben il 22% del Ffo al fondo premiale, l'85% del quale è stato erogato in funzione della Vqr 2011-2014. Sembrerebbe che l'intento del Miur sia stimolare le strutture accademiche a fare ricerca di eccellenza per accedere ai fondi premiali, così da incentivare gli atenei meno produttivi a incrementare la propria *performance* nella ricerca scientifica e colmare il gap con quelli che, invece, hanno già standard elevati. Con queste premesse, la Vqr è lo strumento attraverso il quale indirizzare i flussi di finanziamento pubblici verso i più meritevoli, dando il via a una corsa verso l'eccellenza che, almeno negli auspici del Ministero, dovrebbe motivare tutti gli Atenei ad adottare politiche di miglioramento della propria produzione scientifica.

Ma la relazione virtuosa tra gli esiti della Vqr, redistribuzione meritocratica del fondo premiale e miglioramento generalizzato della qualità della ricerca delle università italiane è un'ipotesi da controllare piuttosto che un assunto da dare per scontato. È infatti possibile che siano per l'appunto i migliori, cioè coloro che già dispongono di risorse

 ISBN 9788835103219

umane e tecniche eccellenti, a vedersi riconosciute quote aggiuntive di risorse, così da consolidare ulteriormente le proprie posizioni di élite nel ranking Anvur. Se così fosse, l'istituzione del fondo premiale starebbe accentuando, e non mitigando, le differenze tra gli Atenei in riferimento alle loro opportunità di accesso alla ricerca di qualità.

Alla luce di queste riflessioni è auspicabile l'avvio di un programma di ricerca di lungo periodo sull'impatto che l'assegnazione dei fondi pubblici secondo una logica premiale sta avendo sulla qualità dei prodotti di ricerca degli atenei, così da rilevare empiricamente quali sono gli effetti di medio e lungo termine che questo tipo di politica sta avendo sull'intero sistema dell'università pubblica in Italia.

Diversi sono gli aspetti che dovrebbero essere presi in considerazione in un disegno di ricerca del genere. In primo luogo, va compreso se le distanze tra i migliori e i peggiori del *ranking*, stilato dall'Anvur in base agli esiti della Vqr 2004-2010 e 2011-2014, si restringono o, al contrario, si dilatano negli esercizi di valutazione a venire. Se le risultanze empiriche confermassero la contrazione di tali distanze, l'impianto generale di valutazione adottato dall'Anvur avrebbe raggiunto gli obiettivi prefissati; nel caso contrario, la logica premiale applicata valutazione degli atenei avrebbe avuto il solo effetto di divaricare ulteriormente le distanze tra i primi del ranking Anvur e chi segue, deprimendo ulteriormente la competitività di questi ultimi.

Il disegno di una ricerca del genere dovrebbe avere una prospettiva longitudinale e tentare di rispondere alle seguenti domande: Come cambiano nel tempo le posizioni degli atenei nel ranking Anvur? Quali sono le Università che hanno consolidato la propria *leadership*? Chi ha migliorato considerevolmente il proprio *score* sugli indicatori di Vqr e chi, invece, non riesce a fare quel cambio di passo che la logica della valutazione premiale prevede?

Una volta individuato "chi" è necessario comprendere "perché". In effetti le variazioni dei punteggi, che nel tempo le strutture riportano sui vari indicatori e indici di Vqr, forniscono solo l'immagine della punta dell'*iceberg* (l'esito della valutazione) mentre il resto (i fattori che concorrono a determinarne l'entità) è sotto la linea dell'acqua. Proviamo a ipotizzare alcune variabili teoricamente correlate con le variazioni agli esiti degli esercizi di valutazione che si succedono nel tempo.

Tra questi vi è senza dubbio la redistribuzione premiale dei fondi pubblici, ma esistono altre variabili di contesto che, a parere di chi

scrive, sono ancor più importanti. Pensiamo al contesto economico del territorio in cui la struttura svolge le proprie attività. È buon senso supporre che elementi quali il Pil regionale, il reddito medio dei nuclei familiari, il livello di occupazione e disoccupazione specialmente giovanile, la presenza/assenza di un tessuto imprenditoriale solido e diffuso sul territorio, la sua vocazione industriale terziaria o agricola, ecc., concorrano a determinare un contesto economico che influenza la quantità di risorse potenzialmente disponibili per la ricerca accademica e i suoi stakeholder.

Anche il contesto sociale struttura le condizioni entro le quali un Ateneo opera. Pensiamo al livello di istruzione degli abitanti del territorio, al loro grado di partecipazione alla vita pubblica della comunità, alla dinamicità della popolazione residente nell'area in riferimento alla sua composizione per genere e per età, alla sua struttura occupazionale/professionale, alle relazioni tra autoctoni e stranieri.

Questi sono solo alcuni esempi di aspetti del contesto socioeconomico che possono essere considerati utili a fare luce sulle differenti opportunità di ricerca su cui gli Atenei possono contare in funzione del loro contesto.

> Le critiche hanno riguardato numerosi aspetti della Vqr (…) Un rilievo particolare dovrebbe essere attribuito al fatto che gli indicatori che ripartiscono la quota premiale trascurano l'influenza di fattori di contesto socio-economico che incidono in modo rilevante sulla capacità di raccogliere tasse e contributi studenteschi, sulla possibilità di attivare risorse esterne, sulla mobilità degli studenti, sulla qualità delle infrastrutture materiali a disposizione, sulla stessa internazionalizzazione. In altre parole, il sistema di ripartizione attuale delle risorse penalizza gli Atenei che operano nei contesti meno sviluppati, alimentando un'assegnazione di risorse che non corrisponde alla qualità della ricerca o della didattica degli Atenei quanto piuttosto al "grado di sviluppo" dei contesti in cui essi operano. Di fatto, più che valutare il merito degli Atenei, i meccanismi premiali attuali finiscono eminentemente per misurare la qualità del contesto. Per questa ragione, la distribuzione attuale delle risorse sta incrementando (e sempre più favorirà, con la progressiva crescita della quota premiale) i processi di divergenza e polarizzazione all'interno del sistema universitario italiano (Realfonzo e Perone, 2016, p. 4).

Si tratta, dunque, di un programma di ricerca che si prefigura di elaborare un modello esplicativo in cui la cui variabile dipendente è una sorta di indice sintetico relativo alla variazione dei punteggi (o della posizione) che una struttura ha totalizzato nel ranking Anvur in

occasione degli esercizi di Vqr che si sono succeduti nel tempo. L'obiettivo è comprendere quali fattori, culturali sociali ed economico-finanziari, incidono maggiormente su tali cambiamenti, e rivalutare alla luce di questa analisi l'effettivo contributo della quota premiale del Ffo.

Spostando l'attenzione dall'impianto generale della Vqr ai suoi aspetti procedurali, nel terzo capitolo Barbanera esamina dettagliatamente le procedure per la formazione dei Gev e della lista dei revisori. La numerosità e la composizione del Gev di Area 14 sollevano non poche perplessità, vista la sproporzione tra l'esiguo numero di Ev che lo compongono e la pluralità di scuole e tradizioni scientifiche cui gli studiosi delle scienze politiche e sociali appartengono. L'Area 14, infatti, è caratterizzata da una notevole frammentarietà disciplinare; lo testimoniano i numerosi Ssd che in essa sono rappresentati. Nel Gev 14, anche solo in considerazione della sua consistenza numerica, rischiano di essere scarsamente o per nulla rappresentate scuole, paradigmi, approcci di ricerca.

Di conseguenza il tema della numerosità/eterogeneità dei componenti dei Gev è rilevante, vista anche la delicatezza del compito assegnato agli Ev nella Vqr 2011-2014: individuare i revisori più adatti alla valutazione dei prodotti conferiti. Ciascun Ev porta nel suo lavoro un *background* di conoscenze e competenze accumulate nel corso della sua carriera, che va necessariamente bilanciato all'interno del Gev di area. Un Gev eterogeneo e in qualche misura rappresentativo delle diverse scuole e settori disciplinari della sua Area di riferimento potrebbe migliorare la gestione della pluralità intra-disciplinare, a partire dalla difficile fase di assegnazione dei prodotti ai revisori. A questo proposito, l'elevato numero di revisioni rifiutate e di giudizi contrastanti può essere considerato un indizio delle difficoltà incontrate dai Gev nel *matching* prodotto-revisore, difficoltà in parte dovute alla elevata pluralità disciplinare dell'Area 14, che ha reso particolarmente problematica l'associazione tra i temi e gli approcci teorico-metodologici di una data pubblicazione e le competenze di un dato revisore.

Tali considerazioni hanno bisogno però di risultanze empiriche a sostegno. A tal proposito, la *Social Network Analysis* (SNA) è un valido strumento di ricerca. Come scrivono Mattioli, Anzera, Toschi (2014, p. 9) «la *Social Network Analysis* (SNA) nasce come tecnica

per determinare e descrivere l'interconnessione delle relazioni tra individui che fanno parte di gruppi, comunità e organizzazioni di varie dimensioni», e dunque potrebbe tornare utile per osservare la composizione della rete dei valutatori realizzata in occasione della Vqr 2011-2014. Nella rete andrebbero inseriti due tipi di attori: i componenti del Gev dell'Area di riferimento, che a questa rete hanno contribuito a dar forma, e i *referees* scelti per valutare le pubblicazioni. Una simile analisi potrebbe servire a individuare, per così dire, la posizione, i confini e la densità interna della rete formata da Ev e revisori, e valutare, in termini anche di rappresentatività, la relazione tra tale spazio e quello costituito dall'ampia ed eterogenea comunità dell'Area 14.

Una *Social Network Analysis* semplificata potrebbe essere ristretta ai soli revisori. Ricostruendo le interazioni reciproche (citazioni e coautoraggi) tra i *referees* della medesima Area Cun, riconoscibili dal Ssd di afferenza, si osserverebbe il grado di omogeneità/eterogeneità della comunità dei pari cui appartengono i revisori che hanno partecipato alla Vqr 2011-2014; tale strategia di analisi sembra particolarmente adatta a quelle aree scientifiche, come l'Area 14, in cui la valutazione è condotta con *peer review* e non tramite bibliometria. In effetti nella Vqr il concetto di pari è un tema delicato. Da un lato, se il gruppo dei pari è molto esteso ed eterogeneo, assisteremmo alla formazione di sottogruppi che aderiscono a paradigmi e scuole differenti, certamente rappresentativi dell'eterogeneità scientifica di cui si è già accennato, ma tra loro difficilmente conciliabili e che, ciò nonostante, sarebbero chiamati a valutarsi reciprocamente. Dall'altro, un gruppo di *referees* che al suo interno è molto omogeneo rischia di trasformare la *peer review* in un esercizio di valutazione incapace di cogliere la qualità nella diversità.

Un altro elemento costitutivo della Vqr è la scheda attraverso la quale i revisori erano chiamati a valutare i prodotti. Nel quarto capitolo Floridi elenca e descrive una serie di criticità: reazione all'oggetto (il revisore assegna un punteggio non alla qualità del prodotto bensì, ad esempio, all'autorevolezza dell'autore che lo ha scritto), domande con oggetti multipli (il revisore è forzato a formulare un giudizio complessivo in riferimento ad aspetti significativamente diversi della qualità scientifica), sotto-determinazione delle domande (alcuni items su cui il revisore è tenuto ad esprimere un giudizio andrebbero articolati in *items* più specifici).

Un altro aspetto controverso, e ampiamente discusso nel capitolo,

riguarda la possibilità concessa al revisore di modificare i punteggi assegnati al prodotto sui tre criteri di qualità dopo aver visionato la classe di merito in cui lo stesso viene collocato al termine della procedura, avendo di fatto la possibilità di riallocare *ex post* il prodotto in una classe di merito diversa da quella che risulta dalla combinazione dei giudizi espressi separatamente per ciascun criterio.

Non sappiamo se l'Anvur abbia a disposizione la tracciatura di tali modifiche. Nel caso fossero disponibili, sarebbe molto interessante confrontare i punteggi assegnati in prima battuta dai revisori ai tre criteri di qualità, la loro eventuale manipolazione a posteriori e i giudizi estesi riportati *a latere*. La ricalibratura *ex post* dei giudizi sui tre criteri di qualità costituisce, a mio avviso, un processo interessante per studiare la relazione tra il concetto di qualità definito operativamente attraverso la scheda Vqr e l'idea di qualità cui, più o meno consapevolmente, ricorrono i revisori nel valutare un prodotto. Infatti, in linea di principio l'esigenza di ritoccare i punteggi attribuiti sulle tre dimensioni una volta aver visto il loro effetto combinato in riferimento alla classe di merito cui il prodotto risulta assegnato è dovuta alla mancata corrispondenza tra la valutazione complessiva della qualità della pubblicazione da parte del revisore e i punteggi da lui stesso espressi sulle tre dimensioni prese separatamente. Un'analisi attenta dei giudizi discorsi espressi dai revisori in queste circostanze può far luce su quali aspetti della qualità di una pubblicazione, eventualmente alternativi o semplicemente diversi rispetto a quelli considerati dall'Anvur, entrano in gioco nell'attività di revisione.

L'analisi dei giudizi estesi potrebbe essere utile anche per stimare l'importanza attribuita dai revisori ai tre criteri nel momento della formulazione del giudizio, fornendo così elementi per una diversa ponderazione dei criteri nella composizione dell'indice numerico e nella conseguente assegnazione a una data classe di merito. In tale maniera «il punteggio assegnato al prodotto su ciascun criterio potrebbe assumere un peso proporzionale alla centralità del criterio stesso rispetto alla determinazione della qualità della ricerca» (Fasanella e di Benedetto, 2015, p. 65).

Oltre che dalla loro idea di qualità e dall'uso conseguente che i revisori fanno della scheda di valutazione, la classe di merito assegnata a una data pubblicazione può dipendere anche da alcune caratteristiche biografiche dei revisori. Potremmo ad esempio controllare l'ipotesi

che esista una sorta di propensione al giudizio più severo tra i revisori con più lunga esperienza accademica (professori associati e ordinari), e che tale propensione venga meno tra i più giovani (i ricercatori).

Altri possibili interrogativi di ricerca sono legati al fatto che la Vqr non è una *blind peer review*, cioè i revisori conoscono i nomi degli autori delle pubblicazioni che valutano. L'attività di valutazione dei revisori più giovani e meno esperti è in qualche modo influenzata dalle informazioni relative al prestigio e all'autorevolezza dell'autore del prodotto o della casa editrice che ha pubblicato il prodotto? C'è la tendenza a valutare meglio alcuni tipi di prodotto (ad esempio gli articoli su riviste prestigiose) a scapito di altri (contributi in volume), a prescindere dal loro contenuto?

Un approccio di ricerca del genere offre notevoli spunti di riflessione, utili a correggere alcune storture insite nella valutazione dei pari. Malgrado i *bias* tipici della *peer review* non possano essere eliminati completamente, un'analisi delle caratteristiche dei revisori, dei prodotti e degli autori, associati ai punteggi relativi agli esiti della valutazione, permetterebbe di individuarli e proporre eventualmente correttivi conseguenti.

Come fisiologicamente avviene in qualsiasi processo di *peer review* basato sulla valutazione congiunta di uno stesso testo da parte di due o più revisori, anche la Vqr ha dato luogo a casi di valutazioni discordanti, cioè situazioni in cui due revisori hanno attribuito una stessa pubblicazione a due diverse classi di merito.

Lo scopo della *peer review* della Vqr è però diverso da quello della *peer review* adottata dalle riviste e dalle collane editoriali. Se nel primo caso l'attribuzione di un giudizio di qualità (graduato lungo una scala) costituisce lo scopo esclusivo della *review*, nel secondo caso il giudizio di qualità è strumentale alla valutazione della pubblicabilità di un saggio ed è solitamente corredato da indicazioni per migliorarne la qualità. In questa seconda situazione l'eventuale discordanza tra due revisioni non deve essere ricomposta attraverso una qualche forma di automatismo (come avviene per la Vqr), quanto piuttosto gestita dall'autore del saggio nel decidere se e come tenere conto delle indicazioni dei revisori per apportare eventuali modifiche al testo.

Le valutazioni discordanti, fisiologiche e ampiamente gestibili nella *peer review* adottata da riviste e collane, sono diventate patologiche nella Vqr per due ordini di motivi: un punteggio medio ottenuto da due

punteggi di partenza distanti tra loro è di per sé meno robusto di un punteggio ottenuto come risultante di punteggi simili, indipendentemente dal criterio che si adotta per costruirlo; la mancata trasparenza – resa a mio avviso ancor più necessaria dal fatto che, come accennato in precedenza, la *peer review* della Vqr non è cieca – nella gestione delle controversie, particolarmente evidente nel caso dell'Area 14, ha contribuito a minare la fiducia della comunità dei valutati nei confronti della Vqr nel suo complesso.

Per questo nell'ambito della Vqr le controversie costituiscono un oggetto strategico di riflessione. Oltre che sulle due classi di merito in cui i revisori hanno classificato un dato prodotto, un'analisi quantitativa, anche a carattere esplorativo, del fenomeno potrebbe basarsi sulle seguenti variabili.

Per i *referees*:

1. il Ssd, in quanto indicatore della competenza in dotazione al revisore del prodotto (è ipotizzabile che i giudizi discordanti su un prodotto possano essere dovuti agli Ssd difformi tra i due revisori);
2. l'età accademica/professionale, in quanto indicatore dell'esperienza maturata in merito alla *peer review*;
3. il ruolo svolto all'interno dell'ateneo/ente di affiliazione (si ritiene che possano essere rilevanti le differenze tra il giudizio di un professore ordinario, rispetto a un associato o un ricercatore);

Per i prodotti si dovrebbe avere a disposizione:

1. le caratteristiche dell'autore o degli autori, tra le quali c'è il Ssd, l'ateneo/ente di affiliazione, la fascia di inquadramento e l'età accademica;
2. il tipo di prodotto, secondo la classificazione fornita dal Miur stesso (a titolo esemplificativo, nell'Area 14 i capitoli di libro hanno ricevuto una percentuale di valutazioni negative maggiore rispetto alle monografie o agli articoli su rivista);
3. la collocazione editoriale.

Dato l'alto numero di variabili che in ipotesi risultano rilevanti, è possibile già da ora parlare di un modello di analisi multivariata dei dati. Potrebbe essere ad esempio interessante capire se il fenomeno sia stato tendenzialmente legato ad alcune caratteristiche dei revisori (età, ruolo accademico, affiliazione, competenze scientifiche, ecc.) oppure se a determinare un alto numero di valutazioni discordanti sia stato il giudizio

difforme su un determinato criterio di qualità, oppure ancora se il problema abbia interessato principalmente una certa categoria di pubblicazioni (articoli su rivista, capitoli, monografie, ecc.) e non altre.

Le possibili linee di ricerca qui delineate non esauriscono certo le potenzialità di studio e di analisi offerte dai dati di processo e di prodotto della Vqr. Esemplificano però come l'accesso da parte della comunità scientifica ad alcune informazioni potrebbe rendere la Vqr una straordinaria opportunità di apprendimento – tanto per chi valuta quanto per chi è valutato, assolvendo così a una delle funzioni principali della valutazione.

 ISBN 9788835103219

 ISBN 9788835103219

Riferimenti bibliografici

Agnoli M.S. (1994), *Concetti e pratica nella ricerca sociale*, FrancoAngeli, Milano.

Aksnes D. W., Langfeldt L., Wouters P. (2019), "Citations, citation indicators, and research quality: an overview of basic concepts and theories", *Sage open*, 9, 1: 1-17.

Anvur (2011), *Valutazione della qualità della ricerca 2004-2010 (Vqr 2004-2010). Bando di partecipazione*, www.anvur.it

Anvur (2015), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Bando di partecipazione*, www.anvur.it

Anvur (2015a), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). La selezione dei componenti Gev*, www.anvur.it

Anvur (2015b), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Linee guida dei gruppi di esperti della valutazione (Gev)*, www.anvur.it

Anvur (2017), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale Anvur. Parte Prima: Statistiche e risultati di compendio*, www.anvur.it

Anvur (2017a), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Politiche e Sociali (Gev 14),* www.anvur.it

Anvur (2017b), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Politiche e Sociali (Gev 14). Appendice A*, www.anvur.it

Anvur (2017c), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Politiche e Sociali (Gev 14). Appendice B*, www.anvur.it

Anvur (2017d), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Politiche e Sociali (Gev 14). Appendice C*, www.anvur.it

Anvur (2017e), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Criteri per la valutazione dei prodotti di ricerca. Appendice A*, www.anvur.it

Anvur (2017f), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area*

 ISBN 9788835103219

Medica (Gev 6). *Appendice*, www.anvur.it
Anvur (2017g), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Agrarie e Veterinarie (Gev 7)*. *Appendice* www.anvur.it
Anvur (2017h), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze dell'antichità, filologico-letterarie e storico-artistiche (Gev 10)*. *Appendice*, www.anvur.it
Anvur (2017i), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Scienze Psicologiche (Gev 11b)*. *Appendice*, www.anvur.it
Anvur (2017l), *Valutazione della Qualità della Ricerca 2011-2014 (Vqr 2011-2014). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area Giuridica (Gev 12)*. *Appendice C*, www.anvur.it
Belleri D. (2014), *Semantic Under-determinacy and Communication*, Palgrave Macmillan, London.
Bence V., Oppenheim C. (2005), "The evolution of the UK's Research Assessment Exercise: publications, performance and perceptions", *Journal of Educational Administration and History*, 37, 2: 137-155.
Bentley R., Blackburn R. (1990), "Changes in Academic Research Performance Over Time: A Study of Institutional Accumulative Advantage", *Research in Higher Education*, 31,4: 327-353.
Bonaccorsi A. (2012), *Potenzialità e limiti della analisi bibliometrica nelle aree umanistiche e sociali. Verso un programma di lavoro*, www.anvur.it
Bornmann L., Daniel H. D. (2008), "What do citation counts measure? A review of studies on citing behaviour", *Journal of Documentation*, 64, 1: 45-80.
Boudon R., Lazarsfeld P.F. (1965), *Introduction,* in Boudon R., Lazarsfeld P.F., eds, *Méthodes de la sociologie: I. Le vocabolaire des sciences sociales*, Mounton & Co., Paris (trad. it.: Cavazzani A.S., a cura di, *L'analisi empirica nelle scienze sociali. Volume I: Dai concetti agli indici empirici*, Il Mulino, Bologna, 1969).
Bressan M. (2007), "Presentazione dei risultati dell'esercizio di valutazione della ricerca 2001-2005", Modena, 14 settembre.
Campelli E. (2001), "Tohu va-vohu. Note non tecniche sul problema della classificazione", *Sociologia e Ricerca Sociale*, 64: 10-21.
Campelli E. (2011), "A proposito di riviste. Valutazione e aritmetica, «protezionismo» e «liberismo»", *Sociologia e Ricerca Sociale*, 32, 95: 5-12.
Cantril H. (1965), *The pattern of human concerns*, Rutgers University Press, New Brunswick.
Chessa S., Vargiu A. (2011), "Valutazione universitaria e mutamenti istituzionali in Europa", *Studi di Sociologia*, 49: 3-34.
Cipriani R. (2013), "È scoppiata la valutazione. Una proposta: il criterio della non prevalenza", *Sociologia e Ricerca Sociale*, 100: 11-16.
Civr (2003a), *Valutazione triennale della ricerca. Bando di partecipazione all'esercizio 2001-2003*, www.civr.it
Civr (2003b), *Linee guida per la valutazione della ricerca*, www.civr.it

Civr (2006), *Vtr 2001-2003. Risultati delle valutazioni di panel di area*, www.civr.it
Coats R., Bordon M., van Leeuwen T. N., van Raan A. (2009), "Scaling rules in the science system: influence of field specific citation characteristics on the impact of individual researcher", *Journal of the America Society for Information Science and Technology*, 60, 4: 740-753.
Collins R. (1994), *Four Sociological Traditions*, Oxford University Press, Oxford (trad. it.: *Quattro tradizioni sociologiche: manuale introduttivo di storia della sociologia*, Zanichelli, Bologna, 1996).
Corbetta P. (2014), *Metodologia e tecniche della ricerca sociale*, Seconda Edizione, Il Mulino, Bologna.
Crane D. (1965), "Scientists at major and minor university. A study of productivity and recognition", *American Sociological Review*, 30, 5: 699-714.
Crane D. (1967), "The gatekeepers of science. Some factors affecting the selection of articles in scientific journals", *American Sociologist*, 2, 4: 195-201.
Decreto Ministeriale n. 458 del 27/06/2015, *Linee guida valutazione qualità della ricerca (VQR) 2011-2014*.
Di Benedetto A. (2015), "Un'analisi del concetto di qualità della ricerca nella Vqr", *Sociologia e Ricerca Sociale*, 108: 95-112.
Di Franco G. (1989), *Qualità della vita: dai modelli alle ricerche empiriche* in Vergati S., a cura di, *Dimensioni sociali e territoriali della qualità della vita*, La Goliardica, Roma.
Enqa (2015), *Standards and guidelines for quality assurance in the European Higher Education Area,* Brussels, Belgium. https://enqa.eu/
European Commission (2010), *Assessing Europe's University Based-Research*, https://ec.europa.eu/info/research-and-innovation
Faggiano M.P. (2012), *Gli usi della tipologia nella ricerca sociale empirica*, FrancoAngeli, Milano.
Fantoni S. (2015), "Il Sistema di valutazione Anvur", *Scuola Democratica*, 3: 695-703.
Fasanella A., Di Benedetto A. (2014), "Luci e ombre nella Vqr 2004-2010. Un focus sulla scheda di valutazione peer nell'Area 14", *Sociologia e Ricerca Sociale*, 104: 59-84.
Fasanella A., Di Benedetto A. (2015), "La valutazione dei pari nelle scienze sociali e politiche. La lezione della Vqr 2004-2010", *Sociologia e Politiche Sociali*, 18, 2: 44-72.
Fasanella A., Martire F. (2017), "Considerazioni metodologiche sulla Vqr 2011-2014 e possibili sviluppi della valutazione", *Sociologia e Ricerca Sociale*, 114: 89-116.
Fideli R. (2001), "La costruzione di un indice tipologico: criteri semantici, numerici ed empirici", *Sociologia e Ricerca Sociale*, 64: 124-137.
Frabboni B., Sacchetta P. (2005), "Vtr 2001-2003: valutazione triennale della ricerca. Struttura del sistema informatico gestito via web", *Cineca Magazine*, 53: 5.
Garfield E. (1979), "Is citation analysis a legitimate evaluation tool?", *Scientometrics*, 1, 4: 359-375.

 ISBN 9788835103219

Geuna A., Martin B.R. (2003), "University Research Evaluation and Funding: an International Comparison", *Minerva*, 41: 277-304.
Glänzel W. (2008), "Seven myths in bibliometrics: about facts and fiction in quantitative science studies", in Kretschmer, H., Havemann F. (eds., 2008), *Proceedings of WIS 2008*, Berlin, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting.
Harvey L. (2008), "Democratizing Quality", Budapest, 20 novembre 2008.
Hceres (2014), *Criteria for the evaluation of research unit: the HCERES standards*, https://www.hceres.fr/en.
Hempel C. G. (1952), *Fundamentals of Concept Formation in Empirical Science*, Chicago, University of Chicago Press (tr. it., *La formazione dei concetti e delle teorie nella scienza empirica*, Feltrinelli, Milano, 1961).
Hicks D. (2011), "Performance based university research funding system", *Research Policy*, 41: 251-261.
Hicks D., Wouters P., Waltman L., de Rijcke S., Rafols I. (2015), "Bibliometrics: The Leiden Manifesto for research metrics", *Nature*, 520, 7548: 429-431.
H.M. Treasury (2006), *Science and Innovation Investment Framework 2004-2014: Next Steps*, www.hm-treasury.gov.uk
Jimenez-Contreras (2010), *Como Utilizar los indacoderes biblimetricos para la evaluacion de la actividad investigadora, la solicitud de sexenios y acreditacion profesores,* www.uc3.ugr.es
Knaw, Vsnu and Now (2009), *Standard Evaluation Protocol 2009-2015: Protocol for Research Assessment in The Netherlands*, www.knaw.nl
Kuhn T.S. (1969), *The Structure of Scientific Revolution*, *Poscritto*, The University of Chicago (trad. it.: Carugo A., a cura di, *La struttura delle rivoluzioni scientifiche*, *Poscritto*, Giulio Enaudi Editore, Torino, 2009).
Lazarsfeld P.F., Barton A.H. (1951), *Qualitative Measurement in the Social Sciences: Classifications, Typologies and Indicies*, in Lerner D. and Lasswell H.D., eds., *The Policy Science*, Stanford University Press, Stanford (trad. it.: *Classificazioni, tipologie e indici*, in Lazarsfeld P. F., *Metodologia e ricerca sociologica*, a cura di Capecchi V., il Mulino, Bologna, 1967).
Lazarsfedl P.F. (1966), *Concept formation and measurement in the behavioral sciences: some historical observation*, in Di Renzo G. J., ed., *Concept, Theory, and Explanation in the behavioral sciences*, Random House, New York (trad. it., *Formazione e misurazione dei concetti nelle scienze del comportamento*, in Lombardo C., a cura di, *Saggi storici e metodologici*, Eucos, Roma, 2001).
Lee C.J., Sugimoto C.R., Zhang G., Cronin B. (2013), "Bias in Peer Review", *Journal of the American Society for Information Science and Technology*, 66, 1: 2-17.
Liani S., Martire F. (2017), *Pretest: Un approccio cognitivo*, FrancoAngeli, Milano.
Mahoney, M.J. (1977), "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System", *Cognitive therapy and research*, 1, 2: 161-175.
Marradi A. (1984), *Concetti e metodi per la ricerca sociale*, La Giuntina, Firenze.
Marradi A. (1990), "Fedeltà di un dato, affidabilità di una definizione operativa", *Rassegna Italiana di Sociologia*, 31, 1: 55-96.

Marradi A. (1994), "Referenti, pensiero e linguaggio: una questione rilevante per gli indicatori", *Sociologia e Ricerca Sociale*, 43: 137-207.
Marradi A. (2007), *Metodologia delle scienze sociali*, Il Mulino, Bologna.
Martini A., Cais G. (2000), *Controllo (di gestione) e valutazione (delle politiche): un (ennesimo ma non ultimo) tentativo di sistematizzazione concettuale*, in Palumbo M., a cura di, *Valutazione 2000. Esperienze e riflessioni, Primo Annuario dell'Associazione Italiana di Valutazione*, FrancoAngeli, Milano.
Matarazzo F. (2018), "Il percorso politico e parlamentare della valutazione nelle università. Una storia lunga quarant'anni", *Articolo* 33, 3: 17-48.
Mattioli F., Anzera G., Toschi L. (2014), *Teoria e ricerca nell'analisi delle reti sociali*, Euroma, Roma.
Mauceri S. (2003), *Per la qualità del dato nella ricerca sociale. Strategie di progettazione e conduzione dell'intervista con questionario*, FrancoAngeli, Milano.
Mauceri S. (2008), "Ri-scoprire l'analisi dei casi devianti. Una strategia metodologica di supporto dei processi teorico-interpretativi nella ricerca sociale di tipo standard", *Sociologia e Ricerca Sociale*, 87: 109-157.
Merton, R.K. (1968a), "The Matthew Effect in Science", *Science*, 159, 3810: 56-63.
Merton, R.K. (1968b), *Social Theory and Social Structure*, Glencoe, Ill (trad. it.: *Teoria e struttura sociale*, Il Mulino, Bologna, 2000).
Mesr (2007), *Les clés de la reforme des universities,* www.nouvelleuniversite.gouv.fr
Mesr (2009), *L'etat de l'enseignement superior et de la recerche en France – 29 indicateurs. Direction de l'evaluation, de la prospective et de la permormance*, Imprimerie Moderne de l'Est.
Minelli E., Rebora G., Turri M. (2008), *The Structure and Significance of the Italian Research Assessment Exercise (VTR)*, in Mazza C., Quattrone P. and Riccaboni A., eds., *European Universities in Transition*. *Issues, Models and Cases.* Cheltenham (UK), Edward Elgar.
Morcellini M. (2015), "Per un'università sostenibile e moderna. Uno sguardo sociologico sulla valutazione e sull'Anvur", *Rassegna Italiana di Valutazione*, 63: 68-82.
Nobile S. (1997), *La credibilità dell'analisi del contenuto*, FrancoAngeli, Milano.
Nobile S. (2008), *La chiusura del cerchio. La costruzione degli indici nella ricerca sociale*, Bonanno Editore, Acireale-Roma.
Osservatorio per la valutazione del sistema universitario (1997), *Ruolo, organizzazione e attività dei Nuclei di valutazione interna delle università, relazione presentata all'"Incontro nazionale sulla valutazione del sistema universitario*, 19 settembre 1997, Doc 5/97.
Otley D. (2010), "Research Assessment in the UK: An Overview of 1992-2008", *Australian Accounting Review*, 20: 3-13.
Palumbo M. (2003), "La valutazione partecipata e i suoi esiti", *Rassegna italiana di valutazione*, 7, 25: 71-88.
Palumbo M., Torriggiani C., a cura di (2009), *La partecipazione fra ricerca e valutazione*, FrancoAngeli, Milano.

Palumbo M., Pennisi C. (2011), "Le ragioni delle regole per la valutazione dell'università: per un'etica della pratica accademica", *Studi di Sociologia*, 49: 35-50.
Palumbo M., Pennisi C. (2014), "La valutazione senza governo", *Rassegna Italiana di Valutazione*, 59: 7-33.
Palumbo R. (2013), *La valutazione periodica della ricerca nelle discipline economico-aziendali*, FrancoAngeli, Milano.
Penfield T., Baker M. J., Scoble R., Wykes M. C. (2014), "Assessment, evaluations, and definitions of research impact: A review", *Research evaluation*, 23, 1: 21-32.
Perotti R. (2008), *L'Università truccata*, Einaudi, Torino.
Pitrone M.C. (2009), *Sondaggi e interviste. Lo studio dell'opinione pubblica nella ricerca sociale*, FrancoAngeli, Milano.
Pitrone M.C., Pavsic R. (2003), *Come conoscere opinioni e atteggiamenti*, Bonanno, Catania.
Polanyi M. (1966), *The Tacit Dimension*, New York, Anchor Books (tr. it. *La conoscenza inespressa*, Armando, Roma, 1979)
Poli S. (2008), *Lo studio degli atteggiamenti nella ricerca sociale: dalle definizioni alle tecniche*, in Bichi R., a cura di, *La distanza sociale. Vecchie e nuove scale di misurazione*, FrancoAngeli, Milano.
Reale E. (2013a), *La valutazione della ricerca pubblica. Un'analisi della valutazione triennale della ricerca,* FrancoAngeli, Milano.
Reale E. (2013b), "La valutazione della ricerca e il cambiamento dell'Università", *Sociologia e Ricerca Sociale*, 100: 148-159.
Reale E., Pennisi C. (2012), "Valutare nella crisi: effetti sull'Università e la ricerca", *Rassegna Italiana di Valutazione*, 14: 7-14.
Realfonzo R., Perone G. (2016), *Qualità degli Atenei e contesto socio-economico. La sperequazione nell'allocazione delle risorse tra le Università italiane*, in Ragozini G., a cura di, *Lo scenario universitario. Mercato del lavoro e sbocchi occupazionali*, FrancoAngeli, Milano.
Rebora G. (2010), *Tra inferno e paradiso. Gli atenei italiani alla prova della valutazione*, Scripta Web, Napoli.
Rebora G. (2013), *Nessuno mi può giudicare: l'università e la valutazione*, Guerini, Milano.
Rebora G., Turri M. (2010), "Lo sviluppo dei sistemi di valutazione della ricerca: un'analisi critica dell'esperienza italiana", Paper presentato al Convegno AIDEA 2010 "Pubblico & non profit per un mercato responsabile e solidale" tenutosi dal 21/10 al 22/10/2010 presso l'Università Bocconi di Milano.
REF (2010), *REF, Impact Pilot Exercise: Findings of the Expert Panels,* http://www.ref.ac.uk
REF (2019), *Panel criteria and working methods*, https://www.ref.ac.uk/
Ribolzi L. (2013), "Valutare l'università: una sfida non solo per l'Anvur", *Sociologia e ricerca sociale*, 100: 23-32.
Rip A., van der Meulen B.J.R. (1995), "The Patchwork of the Dutch Evaluation System", *Research Evaluation*, 5: 45–53.
Rizzi D., Silvestri P. (2002), "La valutazione del sistema universitario italiano: una

storia recente", *Nota di lavoro dell'Università Ca' Foscari di Venezia*, 1: 1-23.
Scarpitti L. (2001), *La valutazione nel sistema universitario italiano*, in Stame N., a cura di, *Valutazione 2001. Lo sviluppo della valutazione in Italia*, FrancoAngeli, Milano.
Schroter S., Black N., Evans S., Carpenter J., Godlee F., Smith R. (2004), "Effects of Training on Quality of Peer Review: Randomised Controlled Trial", *British Medical Journal*, 328, 7441: 673.
Schmitz C. (2008), *Messung Der Forschungsleistung in der Betriebswirtschaftslehre auf Basis der ISI-Zitationsindizes. Eine kritische Analyse anhand konzeptioneller Uber-legungen und empirischer,* Befunde, Eul Verlag.
Searle J.R. (1979), *Expression and Meaning. Studies in the Theory of Speech Acts*, Cambridge, University Press.
Searle J.R. (1980), *The Background of Meaning*, in Searle J.R., Kiefer F., Bierwisch M., eds., *Speech Act Theory and Pragmatic*, Reidel Publishing Company, Dordrecht.
Seglen P.O. (1997), "Citations and journal impact factors: questionable indicators of research quality", *Allergy*, 52, 11: 1050-1056.
Stame N. (2016), *Valutazione pluralista*, FrancoAngeli, Milano.
Travis C. (1975), *Saying and Understanding. A Generative Theory of Illocutions*, Basil Blackwell, Oxford.
Travis C. (1981), *The True and False. The Domain of the Pragmatic*, John Benjamins Publishing Company, Amsterdam-Philadelphia.
Turri M. (2012), "Linee di evoluzione della valutazione nei sistemi universitari europei", *Liuc Papers 259*, *Serie Economia e Impresa*, 67: 1-15.
Tusini S. (2006), *La ricerca come relazione: l'intervista nelle scienze sociali*, FrancoAngeli, Milano.
Valentini E. (2013), "Ritorno al passato? Il cortocircuito riforme/valutazione nel campo delle scienze umanistiche e politiche-sociali", *Sociologia e Ricerca Sociale*, 100: 72-90.
Wennerâs, C., Wold, A. (1999), "Nepotism and Sexism in Peer Review", *Nature*, 387, 6631: 341-343.
Wright Mills C. (1959), *The Sociological Imagination*, Oxford University Press, Oxford (trad. it.: *L'immaginazione sociologica*, Il Saggiatore, Milano, 1962).
Xie Y. (2014), "«Undemocracy»: Inequalities in Science", *Science*, 344, 6186: 809-810.

 ISBN 9788835103219

Curatori e autori

Antonio Fasanella è professore ordinario di Storia e metodo delle scienze sociali, Metodologia della ricerca sociale, Teorie e pratiche della valutazione presso il Dipartimento di Comunicazione e ricerca sociale della Sapienza Università di Roma. Ha pubblicato numerosi saggi, articoli e volumi su temi metodologici e di valutazione sociale.

Fabrizio Martire, professore associato presso il Dipartimento di Comunicazione e ricerca sociale dell'Università di Roma La Sapienza, insegna materie metodologiche e si occupa di temi connessi alla valutazione, alla metodologia della ricerca sociale e alla storia della sociologia.

Lorenzo Barbanera è dottore di ricerca in Comunicazione, ricerca sociale e marketing – con curriculum in Metodologia delle scienze sociali – presso l'Università La Sapienza di Roma. Ha inoltre collaborato con l'istituto Censis, nella divisione Economia e Territorio. Fra i suoi interessi di ricerca vi sono la valutazione delle politiche formative in ambito universitario e la realizzazione di strumenti di rilevazione per indagini standard.

Federica Floridi è dottore di ricerca in Comunicazione, ricerca, innovazione, curriculum in Metodologia delle scienze sociali, presso Sapienza Università di Roma. Collabora con enti pubblici e privati di valutazione ed è stata responsabile per gli aspetti tecnico-metodologici di ricerche valutative nell'ambito di progetti e programmi della coo-

 ISBN 9788835103219

perazione internazionale del Ministero degli Affari esteri. I suoi principali interessi di ricerca sono la valutazione delle politiche pubbliche e sociali, i processi di scolarizzazione, la valutazione dell'università e della ricerca.

Federica Fusillo ha conseguito il dottorato di ricerca in Comunicazione, ricerca, innovazione, curriculum in Metodologia delle scienze sociali, presso Sapienza Università di Roma; ha conseguito la laurea magistrale presso lo stesso Ateneo con una tesi sull'Abilitazione Scientifica Nazionale. L'interesse verso la valutazione delle politiche formative e, in particolare, del sistema universitario italiano è proseguito nel corso degli anni con la partecipazione a indagini e progetti di ricerca sulla valutazione della qualità della ricerca, sulla produttività scientifica dei docenti universitari, sulle carriere accademiche e sui criteri di valutazione Asn.

Marco Palmieri è attualmente assegnista di ricerca per il Dipartimento di Studi politici e sociali dell'Università di Salerno. È docente a contratto di Metodologia della ricerca sociale presso il Dipartimento di Scienze umane dell'Università dell'Aquila. È anche docente al Master interuniversitario "Sociologia: Teoria, Metodologia, Ricerca" (So.Te.Me.Ri.), organizzato da Università Sapienza, Università di Tor Vergata e Università di Roma 3. Tra i suoi principali interessi di ricerca vi è la costruzione di strumenti standardizzati per lo studio di atteggiamenti, opinioni e valori.

 ISBN 9788835103219

A cura di Antonio Fasanella e Fabrizio Martire

Valutazione della ricerca e ricerca sulla valutazione

Il volume propone una riflessione critica sul tema della valutazione della qualità della ricerca condotta entro una prospettiva orientata in senso progressivo. Non si intende negare l'utilità o, se si vuole, la necessità di sottoporre a valutazione gli esiti della ricerca, ma si ribadisce altresì l'indispensabilità di un controllo pubblico dei processi di valutazione, perseguendo inflessibilmente un principio di *accountability*, al fine di superare la falsa contrapposizione tra approcci *judgemental* e approcci standardizzati alla valutazione.
Se l'unica scelta praticabile è tra una valutazione basata esclusivamente su conoscenza tacita non ricostruibile, su risorse squisitamente soggettive come l'intuizione, e una valutazione ispezionabile/replicabile perché basata su indicatori quantificabili, la tentazione della bibliometria e dell'uso incondizionato dei ranking delle riviste diventa irresistibile. Tutta la trattazione qui svolta sostiene fermamente la possibilità e l'opportunità di una soluzione alternativa: un sistema di valutazione incentrato sulla *peer review* e che rispetti, in ogni suo passaggio, il principio dell'*accountability*. Il volume prende atto positivamente degli sforzi che sono stati condotti in sede Vqr in questo senso, ma non può non riconoscere, d'altra parte, che l'attuale Vqr, nonostante tali sforzi, è ancora contraddistinta da elementi di opacità, di ambiguità, di arbitrio, di scarsa ispezionabilità delle procedure, che la rendono suscettibile di essere perfezionata, anche sulla base delle proposte qui avanzate.

Antonio Fasanella è professore ordinario di Storia e metodo delle scienze sociali, Metodologia della ricerca sociale, Teorie e pratiche della valutazione presso il Dipartimento di Comunicazione e ricerca sociale della Sapienza Università di Roma. Ha pubblicato numerosi saggi, articoli e volumi su temi metodologici e di valutazione sociale.

Fabrizio Martire, professore associato presso il Dipartimento di Comunicazione e ricerca sociale della Sapienza Università di Roma, insegna materie metodologiche e si occupa di temi connessi alla valutazione, alla metodologia della ricerca sociale e alla storia della sociologia.

FrancoAngeli
La passione per le conoscenze