SAPIENZA

UNIVERSITÀ DI ROMA

DIPARTIMENTO DI COMUNICAZIONE E RICERCA SOCIALE
DOTTORATO DI RICERCA IN METODOLOGIA DELLE SCIENZE SOCIALI (MESS)
Ciclo XXVII

Un'analisi metodologica della valutazione dei prodotti nella VQR 2004-2010

Annalisa Di Benedetto

Tutor: Enzo Campelli, Antonio Fasanella

Indice

Introduzione

La scelta della valutazione della ricerca come oggetto d'indagine è certamente connesso alla sua attualità nel panorama italiano, ma deriva principalmente dall'interesse per l'acceso dibattito metodologico intorno ai criteri e agli strumenti in uso in questo campo.

L'esercizio di valutazione della ricerca più esteso mai portato a termine in Italia è l'occasione ideale per condurre un'accurata analisi metodologica, anche alla luce dell'ampia letteratura disponibile sulla valutazione della ricerca. Se è vero che qualunque indagine deve rispondere della propria adeguatezza metodologica per potersi definire scientifica, le indagini valutative devono farlo anche per una questione di responsabilità: i loro risultati dovrebbero, infatti, guidare, o almeno orientare, decisioni concrete. Questa esigenza di *accountability* (Palumbo, 2001) implica per le indagini valutative non solo la rendicontazione completa delle scelte operate, delle attività condotte e dei risultati ottenuti, ma anche la loro validazione tramite fasi di partecipazione, negoziazione e condivisione con gli attori coinvolti (Fasanella, 2013).

L'analisi condotta offre diversi spunti di riflessione e tocca problematiche centrali per la metodologia delle scienze sociali. Pur partendo da una ricostruzione storica delle esperienze di valutazione della ricerca in Italia, l'obiettivo essenziale del lavoro è l'analisi dell'esercizio di valutazione VQR 2004-2010. Si mira, dunque, all'inquadramento dell'esercizio di valutazione dal punto di vista delle politiche, degli approcci e delle tecniche valutative utilizzate, ma si intende soprattutto realizzare un'approfondita analisi metodologica delle procedure relative alla valutazione dei prodotti della ricerca. Il materiale su cui l'analisi si basa è costituito, essenzialmente, dall'insieme dei documenti ufficiali pubblicati circa l'esercizio: il bando, il rapporto finale dell'Agenzia, i rapporti finali di Area e i relativi allegati, cui si aggiungono alcune interviste focalizzate rivolte a testimoni privilegiati e realizzate al fine di chiarire alcuni punti delle procedure, rilevando allo stesso tempo le loro opinioni.

La rendicontazione del lavoro di tesi è organizzata in tre capitoli principali riguardanti i tre aspetti essenziali dell'analisi metodologica: la concettualizzazione della qualità della ricerca, la sua definizione operativa e la rilevazione dei dati. A chiusura dell'analisi metodologica un capitolo è dedicato all'esposizione di alcune proposte per la messa a punto di procedure alternative, più attente alla qualità dei dati riferibili alla qualità dei prodotti della ricerca. I capitoli fondamentali sono preceduti da due capitoli introduttivi: uno dedicato alla ricostruzione delle diverse esperienze di valutazione della ricerca in Italia, l'altro mirato all'esposizione degli obiettivi dell'analisi e degli strumenti utilizzati. Infine l'ultimo capitolo è riferito agli obiettivi e ai possibili impatti, più o meno attesi e desiderati, dell'esercizio di valutazione.

Capitolo 1
La valutazione della ricerca in Italia

Introduzione

La valutazione è stata introdotta nel sistema universitario italiano, almeno sulla carta, nei primi anni ’90, tuttavia solo a partire dalla fine del decennio, a seguito di una serie di riforme, ha assunto un ruolo attivo, seppure reso precario proprio dall’incertezza del quadro normativo mirato a potenziarla. Fino agli anni ’90 dunque le università, e gli enti di ricerca pubblici, in Italia hanno svolto le proprie attività senza doverne rendere conto formalmente all’esterno.

In un sistema di *policy* sostanzialmente universalistico la spesa per la ricerca era giustificata dal suo fine prima ancora che dai suoi risultati. Questa visione inizierà a incrinarsi negli anni ’80 con la diffusione in Europa di politiche di stampo thatcheriano e con l’affermarsi della richiesta di risultati adeguati alle spese. In questo periodo le questioni dell’*accountability*, dell’efficienza e dell’efficacia della spesa pubblica investono non solo il welfare, ma anche il sistema formativo e quello della ricerca. La valutazione della ricerca scientifica viene allora introdotta in diversi Pesi europei, prima nel Regno Unito con un sistema di valutazione completo, il *Research Assessment Exercise*, nel 1986, poi in altri Stati, per lo più nel quadro della pianificazione, come in Germania, Irlanda, Svezia e Finlandia. In diversi casi, come in quello dell’Italia e della Spagna, il principio della valutazione viene introdotto a livello normativo tra la fine degli anni ’80 e l’inizio dei ‘90, ma subisce una lunga fase di gestazione in cui le attività sono limitate e vengono più volte ridefinite.

L’intento di questo capitolo è duplice, da un lato, infatti, è necessario delineare il contesto normativo e accademico in cui l’esercizio di valutazione della ricerca che si intende analizzare è stato implementato, dall’altro risulta fondamentale una ricostruzione delle esperienze e delle pratiche su cui questo esercizio si innesta per poter meglio comprendere il suo impianto generale, alcune delle sue principali caratteristiche, nonché le reazioni che ha suscitato dentro e fuori il mondo accademico.

1.1 L’introduzione della valutazione nell’università italiana

La valutazione è stata introdotta nell’università italiana nei primi anni ’90, a seguito dell’impostazione del principio di autonomia in ambito universitario della legge Ruberti (169/1989). Si assumeva era che gli atenei, resi più autonomi dal punto di vista decisionale e gestionale, dovessero rendere conto dei propri risultati tanto dal punto di vista della didattica quanto da quello della ricerca. La legge 537 del 1993 stabiliva, infatti, un nesso tra il principio di autonomia e quello della valutazione, e prevedeva delle ricadute degli esiti delle valutazioni sull’assegnazione delle risorse. La stessa legge istituiva inoltre i nuclei di valutazione interna «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 e il buon andamento dell’azione

amministrativa» (art. 5) e prevedeva l'istituzione di un Osservatorio Nazionale con il compito di effettuare la valutazione «dell'efficienza e della produttività delle attività di ricerca e di formazione» sulla base delle relazioni dei nuclei di valutazione interna.

Fatta eccezione per alcune "esperienze pilota", legate più ancora che al quadro legislativo al crescente interesse per il tema della valutazione (Rebora, 2012), questi principi resteranno essenzialmente senza applicazione, nonostante l'istituzione dei nuclei di valutazione d'Ateneo e dell'Osservatorio per la Valutazione del Sistema Universitario (OSVU). La stessa creazione dell'Osservatorio avviene nel 1996[1], diversi anni dopo la decisione di istituirlo.

In questa fase l'introduzione della valutazione nel sistema universitario italiano appare puramente formale e leggendo i documenti prodotti dall'Osservatorio si ha l'impressione che i singoli nuclei non si spingano oltre la semplice raccolta dei dati[2]. Nessun esercizio di valutazione esterna è stato sviluppato, nonostante la proposta di un programma di Valutazione Istituzionale dell'Università (VIU) (OVSU, 1999a).

Parallelamente al programma VIU sul fronte della valutazione della ricerca era stato inoltre proposto un programma di Valutazione della Produzione Scientifica delle università (VPS) (OVSU, 1999b), ispirato al *Research Assessment Exercise* (RAE) inglese, principalmente centrato sulla quantità della produzione e mirato alla valutazione degli atenei. Neppure questo progetto troverà, però un'applicazione, e sarà abbandonato a seguito di un riassetto del sistema della valutazione.

Nella seconda metà degli anni '90 erano già attive agenzie di valutazione dell'università e della ricerca in diversi paesi europei (ad esempio la *Quality Assurance Agency for Higher Education* (QAA) in Inghilterra, il Consiglio finlandese per la valutazione dell'istruzione superiore (FHEEC) in Finlandia, l'Agenzia Nazionale per l'Istruzione Superiore in Svezia, e il Consiglio Universitario (CU) in Spagna), e dove queste mancavano erano le conferenze dei rettori (ad esempio in Germania, Svizzera e Olanda) o le associazioni di atenei (ad esempio in Portogallo) a occuparsi delle attività di valutazione[3]. Era frequente in questo periodo il ricorso a una valutazione in due fasi, la prima di auto-valutazione, la seconda esterna, la stessa procedura che nelle intenzioni doveva essere implementata dall'OVSU.

L'architettura complessiva del sistema di valutazione nel sistema universitario italiano venne ridefinita nel 1999 con la legge 370. Questa ridefinizione era legata tanto alla riforma del "3+2", il D.M. 509/1999 con cui l'Italia applicava le linee guida del processo di Bologna, quanto alla legge 59/1997, che indicava nella valutazione dei risultati uno dei punti chiave del governo nazionale della ricerca (art. 20, comma 1), accanto al coordinamento e alla programmazione e gestione delle risorse[4].

---

[1] L'OVSU fu istituito con un decreto del MURST il 22 febbraio 1996.

[2] I nuclei sono organi interni agli atenei, nel 1996 l'80% dei loro membri era costituito da personale universitario (OSVU, 1997b). Rizzi e Silvestri (2002) ricordano che anche se nel 1996 su 56 atenei pubblici 51 avevano istituito un nucleo di valutazione interna, solo 35 di essi avevano steso effettivamente la relazione annuale e solo 25 l'avevano inviata al MURST come previsto dalla normativa (OSVU, 1997b).

[3] Facciamo qui riferimento alla panoramica fornita nel rapporto *Evaluation of European higher education: a status report* redatto dal Centre for Quality Assurance and Evaluation of Higher Education (Denmark) e dal Comité National d'Evaluation (France) per la Commissione Europea nel 1998.

4 L'implementazione della valutazione del sistema universitario risultava infatti pienamente in linea con il contenuto della "Bozza Martinotti", che rilevava come all'aumento dell'autonomia dovesse corrispondere un rafforzamento della funzione di governo, dunque delle capacità di conoscenza, indirizzo, coordinamento e verifica dei risultati da collegare a un sistema di incentivi basati, appunto, sulla valutazione (Martinotti, 1997).

Il Comitato di Indirizzo per la Valutazione della Ricerca (CIVR) era già stato istituito nel 1998[5], con il compito di svolgere l'attività di valutazione al fine di promuovere la qualità della ricerca scientifica e tecnologica nazionale. Con riferimento alla didattica universitaria era invece la stessa legge 370/1999 a istituire il Comitato Nazionale per la Valutazione del Sistema Universitario (CNVSU), e ampliare le funzioni dei nuclei di valutazione[6].

Questi due organi dedicati alla valutazione del sistema universitario hanno operato però in una condizione di incertezza normativa e sostanzialmente come enti depotenziati. Il CNVSU ha condotto un'attività di monitoraggio della didattica, pubblicando un rapporto annuale dal 2000 al 2010, mentre il CIVR ha svolto un solo esercizio di valutazione, la Valutazione Triennale della Ricerca (VTR) 2001-2003. Sul fronte della ricerca il CIVR era stato incaricato di un secondo esercizio di valutazione: la Valutazione Quadriennale della Ricerca (VQR 2004-2008), che non è stato mai realizzato mentre l'acronimo è stato adottato, con un diverso significato, per la Valutazione della Qualità della Ricerca (VQR) svolta sul settennio 2004-2010.

L'Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (Anvur), destinata a sostituire tanto il CIVR quanto il CNVSU, viene istituita già con la legge 286 del 2006 «al fine di razionalizzare il sistema di valutazione della qualità delle università e degli enti di ricerca pubblici e privati destinatari di finanziamenti pubblici, nonché dell'efficienza e dell'efficacia dei programmi statali di finanziamento e di incentivazione delle attività di ricerca e di innovazione» (art. 2, comma 138).

L'istituzione dell'Anvur, concepita come agenzia unica sul modello di analoghe strutture in altri paesi europei[7] e conforme ai criteri ENQA[8], risulta tanto più importante in quanto i risultati delle sue attività «costituiscono criterio di riferimento per l'allocazione dei finanziamenti statali alle università e agli enti di ricerca» (legge 286/2006, art. 2, comma 139).

La sostituzione di CIVR e CNVSU da parte dell'Anvur non è stata però immediata, e il nuovo ente è stato regolamentato solo nel 2010[9], a quattro anni dalla sua istituzione, per divenire operativo nella primavera del 2011, dopo la nomina dei componenti del comitato direttivo.

In questo quadro complessivo vengono più volte riviste le norme per il reclutamento di docenti e ricercatori[10]. L'obiettivo sostanziale del riordino del reclutamento è di permettere agli atenei di

[5] Con il decreto legislativo 204 del 5 giugno 1998, a seguito della delega assegnata al governo proprio dalla legge 59/1997.

[6] All'articolo 1 si legge: «le università adottano un sistema di valutazione interna della gestione amministrativa, delle attività didattiche e di ricerca, degli interventi di sostegno al diritto allo studio, verificando, anche mediante analisi comparative dei costi e dei rendimenti, il corretto utilizzo delle risorse pubbliche, la produttività della ricerca e della didattica, nonché l'imparzialità e il buon andamento dell'azione amministrativa» (legge 370/1999 art. 1).

[7] In particolare sulla scia del modello francese dell'*Agence d'évaluation de la recherche et de l'enseignement supérieur* (AERES), istituita proprio nel 2006 dalla legge di programmazione della ricerca, operativa già nel marzo 2007.

[8] Oltre all'uso di procedure di *quality assurance* esterna, l'indipendenza, la disponibilità di risorse umane e finanziarie adeguate agli scopi. L'ENQA è sostanzialmente un'agenzia europea dedicata alla *quality assurance* delle agenzie di valutazione e ha lo scopo di uniformare i criteri e le metodologie valutative, nel quadro del processo di Bologna.

[9] Con il DPR n. 76 del 1 Febbraio 2010.

[10] La legge 210 del 3 Luglio 1998 istituisce le Valutazioni Comparative per il reclutamento dei professori e dei ricercatori universitari; la legge 230/2005 prevede nuove disposizioni e delega al governo il riordino del reclutamento dei professori universitari, e a questa seguono il d.l. 164/2006 recante il "Riordino della

selezionare i propri docenti e ricercatori in linea con il principio di autonomia, a seguito però di una doppia valutazione: quella a livello nazionale, che conferisce l'abilitazione scientifica ai candidati o l'idoneità (ex. legge 180/2008), e quella svolta dall'ateneo in base al proprio regolamento. Ultima in ordine di tempo, la legge 240/2010, reca ulteriori "Norme in materia di personale accademico e reclutamento, nonché delega al governo per incentivare la qualità e l'efficienza del sistema universitario". L'avvio dell'attività dell'Anvur ha dunque coinciso con l'attuazione dell'ultima riforma, nota come Legge Gelmini, che ha fondato il governo del sistema universitario e delle strutture proprio sulla valutazione, tanto con riferimento ai finanziamenti quanto con riferimento al reclutamento del personale accademico.

Già dall'inizio degli anni 2000 la valutazione era uscita dall'ambito puramente normativo, ma in tempi recenti è entrata nella quotidianità delle istituzioni accademiche; non si tratta di una pratica che riguarda esclusivamente i componenti dei nuclei di valutazione e i responsabili delle strutture, ma di un'attività che coinvolge, e interessa direttamente, i singoli docenti e ricercatori. Questo è uno dei motivi fondamentali per cui il dibattito sulla valutazione dell'università e della ricerca in particolare è diventato non solo più serrato ma anche, forse soprattutto, più diffuso e transdisciplinare, non di rado incentrato proprio sulle procedure adottate per condurre le attività di valutazione.

1.2 Il Comitato di Indirizzo per la Valutazione della Ricerca (CIVR) e la valutazione triennale della ricerca VTR 2001/2003

Il Comitato di Indirizzo per la Valutazione della Ricerca istituito con la legge 204/1998 aveva il compito di operare «per il sostegno alla qualità e alla migliore utilizzazione della ricerca scientifica e tecnologica nazionale, secondo autonome determinazioni con il compito di indicare i criteri generali per le attività di valutazione dei risultati della ricerca, di promuovere la sperimentazione, l'applicazione e la diffusione di metodologie, tecniche e pratiche di valutazione, degli enti e delle istituzioni scientifiche e di ricerca, dei programmi e progetti scientifici e tecnologici e delle attività di ricerca, favorendo al riguardo al confronto e la cooperazione tra le diverse istituzioni operanti nel settore, nazionali e internazionali» (art. 5, comma 1)[11].

Il primo mandato ricevuto dal CIVR per la valutazione della ricerca, risalente al 2002, prevedeva la stesura di linee guida per la realizzazione di un processo di valutazione degli istituti di ricerca e dei progetti speciali finanziati dal MIUR nei tre anni precedenti (Reale, 2008). Le linee guida provvisorie elaborate dal CIVR sono state presentate alla comunità scientifica nel febbraio del 2003,

---

disciplina del reclutamento dei professori universitari, a norma dell'art. 1, comma 5 della legge 4 novembre 2005, n. 230", il d.l. 180/2008 recante "Disposizioni urgenti per il diritto allo studio, la valorizzazione del merito e la qualità del sistema universitario e della ricerca" .

[11] Il Comitato doveva essere composto «da non più di sette membri, anche stranieri, di comprovata qualificazione ed esperienza, scelti in una pluralità di ambiti metodologici e disciplinari» (art. 5, comma 1), e durare in carica per quattro anni (D.P.C.M. 10273 del 26/03/1999). Tra i compiti del CIVR figuravano la predisposizione di rapporti annuali da trasmettere al Ministero dell'Università, al Comitato interministeriale per la programmazione economica (CIPE) o ad altri ministeri interessati, la valutazione degli interventi statali per la ricerca applicata e delle attività svolte per il raggiungimento degli obiettivi del Piano Nazionale per la Ricerca (PNR).

e gruppi di lavoro formati da rappresentanti di CUN, CRUI, singoli enti di ricerca, Confindustria e Confcommercio hanno contribuito alla loro messa a punto[12].

Il decreto ministeriale che indiceva l'esercizio indicava il periodo di riferimento (il triennio 2001-2003) e le strutture cui l'esercizio sarebbe stato rivolto:

«a) Università (statali, non statali legalmente riconosciute);

b) Enti di ricerca di cui all'art. 8 del D.P.C.M. 30.12.1993, n. 593 e successive modificazioni e integrazioni, ENEA e ASI (di seguito denominati enti di ricerca);

c) altri soggetti pubblici e privati che svolgono attività di ricerca, su esplicita richiesta, previa intesa nella quale sia espressa anche la determinazione delle risorse, a carico degli stessi, da destinare al processo di valutazione» (DM 2206 del 16 dicembre 2006, art. 1) [13].

Lo stesso decreto stabiliva criteri di valutazione, procedure e tempi di realizzazione in conformità con le linee guida proposte dal Comitato.

Nelle linee guida la valutazione che interessava la produzione scientifica era definita retrospettiva, mirata cioè a rilevare i risultati tecnico scientifici della ricerca e le loro ricadute socio-economiche, tenendo conto della produttività, della qualità e della rilevanza dei risultati, del confronto tra i risultati previsti e quelli raggiunti, dell'*outcome* della ricerca, il suo impatto socio economico e della capacità di gestione delle risorse della struttura (CIVR 2003)[14].

Il processo di valutazione (Figura 1) prevedeva che le strutture inviassero al CIVR dei dati di contesto attraverso le relazioni dei nuclei di valutazione interna e una selezione di prodotti della ricerca ai *panel* di esperti per la loro valutazione. I *panel* avevano il compito di sottoporre alla peer review i prodotti e i progetti, ottenendo su ciascuno un giudizio di merito, e di redigere un rapporto finale con cui rendicontare i risultati ottenuti alla comunità scientifica e la realizzazione di un *ranking* di area con l'assegnazione delle strutture a fasce di merito predefinite dal CIVR. Infine il Comitato, ricevuti i giudizi di merito, i rapporti di area e i dati di contesto, aveva il compito di produrre dei rapporti finali riferiti alle singole strutture, esprimendo un giudizio di merito per ciascuna di esse.

Sostanzialmente la valutazione avveniva a due livelli: quello delle strutture e quello dei panel di area. Al primo livello di valutazione le strutture partecipanti erano impegnate a selezionare autonomamente un numero di prodotti della ricerca pari alla metà del numero di ricercatori equivalenti a tempo pieno (ETP) impiegati nella struttura stessa[15] (CIVR, 2006a). Le stesse strutture indicavano l'area disciplinare in cui collocare il prodotto, indipendentemente dal settore disciplinare dell'autore e senza alcun obbligo di coprire tutte le aree[16].

[12] La versione definitiva delle linee guida per la VTR è del maggio 2003 (CIVR, 2003), ma il mandato formale per lo svolgimento dell'esercizio è conferito al CIVR nel 2006 (con il DM 2206 del 16 dicembre 2006).

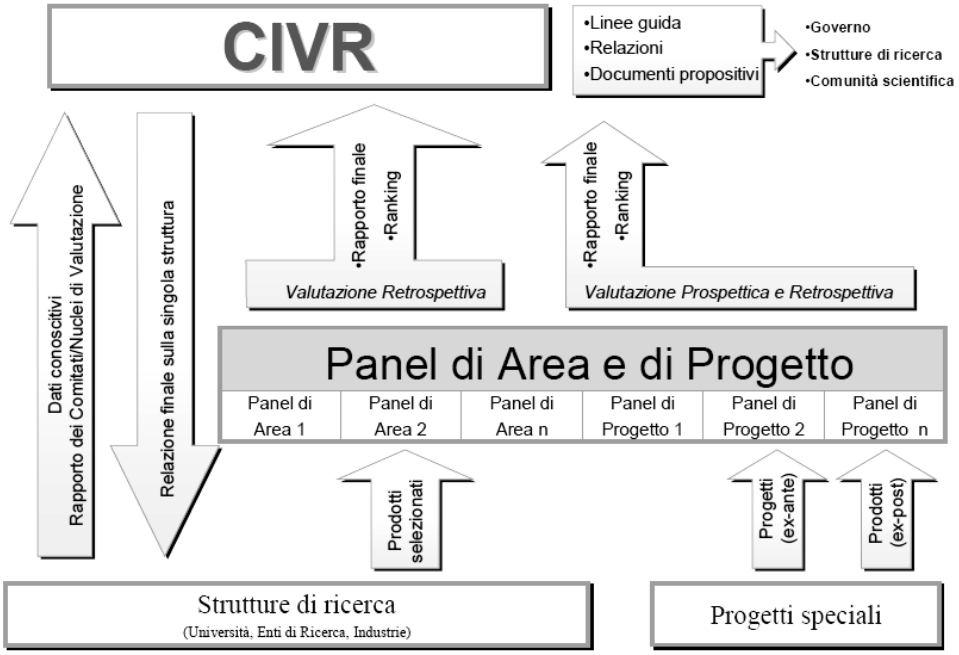
[13] Va sottolineato che la VTR non interessava soltanto le strutture, ma anche i progetti speciali di ricerca previsti e finanziati dal PNR a cui venivano assimilate le attività svolte in regime di agenzia da Enti di ricerca controllati dal MIUR (DM 2206 del 16 dicembre 2006 art. 1, commi 3 e 4).

[14] In riferimento ai progetti speciali alla valutazione retrospettiva si affiancava una valutazione prospettica da condurre prima e durante la realizzazione del progetto in riferimento alle proposte, agli obiettivi, alla qualità e alla rilevanza del progetto, ma anche alle sue potenzialità dal punto di vista della produttività e dell'impatto economico e sociale (CIVR 2003).

[15] Per le Università ciascun ricercatore veniva assimilato a 0.5 ETP, in ragione del doppio compito istituzionale della struttura: ricerca e formazione.

[16] Infatti ai fini della VTR le aree per le quali non erano stati presentati prodotti sono state considerate non attive.

Figura 1 – Organizzazione generale del processo di valutazione nella VTR 2001-2003 (CIVR, 2003, p.16)



Al secondo livello di valutazione si trovavano i panel di esperti, uno per ciascun settore scientifico-disciplinare del CUN[17] più sei per le Aree Speciali, individuate dal CIVR sulla base delle indicazioni del Programma Nazionale della Ricerca e dei programmi di ricerca e sviluppo e multidisciplinari. I panel, composti da un minimo di cinque a un massimo di nove membri in ragione della complessità della disciplina e del numero di prodotti ricevuti, avevano il compito di ottenere dei giudizi di merito su ciascun prodotto avvalendosi dell'opinione di almeno due esperti. Secondo le indicazioni del CIVR (2003), i panel dovevano risultare composti da membri con le più varie conoscenze e competenze, provenienti tanto dall'Università quanto da altri enti e dal mondo produttivo. Il codice di condotta cui questi membri dovevano attenersi imponeva di operare come soggetti indipendenti e non rappresentativi delle istituzioni di afferenza, di assicurare la continuità nella partecipazione ai lavori e la riservatezza, di stabilire anticipatamente i potenziali casi di conflitto di interesse in relazione ai prodotti sottoposti a valutazione (Reale, 2008).

I *panel* avevano la responsabilità diretta della valutazione dei prodotti, e operavano autonomamente nel quadro delle regole fissate dal CIVR. Reale (2008) tuttavia rileva come nei rapporti finali pubblicati dal CIVR si noti l'assenza di riferimenti ai criteri e alle norme che hanno determinato la selezione degli esperti esterni che hanno valutato i prodotti.

Agli esperti era richiesto un giudizio descrittivo (in 500 caratteri), per ciascuno dei criteri individuati dal CIVR:

a) *qualità*: «posizionamento del prodotto rispetto all'eccellenza scientifica nella scala di valore condivisa dalla comunità scientifica internazionale» (CIVR, 2006a, p. 798);
b) *rilevanza*: «valore aggiunto per l'avanzamento della conoscenza nel settore e per la scienza in generale, nonché per i benefici sociali derivati, anche in termini di

[17] Il DM 2206 prevedeva la possibilità di costituire sub-panel, in relazione a specifiche esigenze numeriche o disciplinari presentate dai Panel; in nessun caso però nel corso della VTR i *panel* hanno proposto la costituzione di *sub panel* (CIVR 2006a).

appropriatezza, efficacia, tempestività e durata delle ricadute; integra il giudizio di qualità» (*ibidem*);

c) *originalità/innovazione*: «contributo a nuove acquisizioni o all'avanzamento di conoscenze, nel settore di riferimento; integra il giudizio di qualità» (*ibidem*);
d) *internazionalizzazione*: «posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica, inclusa la collaborazione esplicita con ricercatori e gruppi di ricerca di altre nazioni» (*ibidem*).

In aggiunta ai commenti descrittivi gli esperti fornivano un giudizio complessivo con l'attribuzione di ciascun prodotto a un livello di merito (eccellente, buono, accettabile, limitato). Trattandosi di un procedimento di *informed* peer review i *referee* avevano a propria disposizione, oltre che il prodotto in sé, alcuni elementi aggiuntivi che potevano essere utilizzati per la formulazione del giudizio: una scheda descrittiva del prodotto e, nel caso di articoli, dell'*impact factor* della rivista.

La valutazione finale dei prodotti era fondata sul riesame critico, da parte dei panel, dei giudizi espressi dagli esperti, e sulla formulazione di un giudizio di sintesi, in termini di livello di merito, assunto a maggioranza (CIVR, 2004).

Il grado di concordanza dei giudizi nel corso della VTR, che può essere assunto come un indicatore dell'affidabilità del processo di peer review, è stato analizzato e giudicato positivamente: «poiché le discrepanze tra i giudizi degli esperti non sono marcate esse possono essere riassorbite nel giudizio finale del *panel*» (Reale, 2008, p. 160).

La valutazione dei prodotti della ricerca del triennio 2001-2003 costituiva solo una parte, per quanto importante, dell'intero esercizio di valutazione. Il modello di valutazione includeva, oltre ai i risultati della valutazione dei pari della produzione scientifica, una serie di indicatori volti a rilevare altri aspetti, in particolare: la capacità di sostenere la ricerca nel medio periodo e le sue ricadute applicative (CIVR 2003). La valutazione era dunque basata sulla *performance* scientifica e su dati contestuali, pur attribuendo un peso decisamente maggiore alla prima (Tabella 1).

Tabella 1 – I criteri di valutazione della VTR (Reale, 2008, p. 19)[18]

Valutazione della produzione scientifica del triennio 2001-2003	Qualità, rilevanza, originalità, internazionalizzazione dei prodotti presentati	4.0
	Grado di proprietà dei prodotti eccellenti	2.0
Capacità di sostenere la ricerca nel medio periodo	Ricercatori in formazione	0.5
	Mobilità internazionale	1.0
	Capacità di attrarre risorse per la ricerca dall'esterno	1.0
	Investimenti di risorse proprie per la ricerca	0.5
Ricadute applicative	Valorizzazioni applicative	0.5

In sintesi l'esercizio mirava a rilevare non la produttività, ma la qualità della produzione scientifica delle strutture. Il modello è stato costruito per valutare le strutture a livello di area disciplinare, non i singoli ricercatori, considerando un numero contenuto di prodotti, quelli individuati come eccellenti. Va rilevato che, considerando solo parte della produzione scientifica

[18] Nella tabella non viene riportato uno degli indicatori "Capacità di impegnare risorse umane", per cui era previsto un peso di 0,5, ma che non è stato impiegato ai fini della VQT dato che: «l'eterogeneità dei dati disponibili nella banca dati MIUR ne sconsiglia l'impiego» (Civr, 2006a, p. 617).

nazionale, la VTR non era in grado di fornire una valutazione sull'intero sistema di ricerca pubblica italiano, evitava dunque qualsiasi rischio relativo all'adozione di una visione *publish or perish* della ricerca. Da un punto di vista differente però le regole alla base del bando VTR conferivano forse un'eccessiva libertà alle *strutture* che, per assurdo, avrebbero potuto inviare a valutazione i prodotti di un solo autore, per una sola area.

Reale (2008) a conclusione del suo esame dei risultati della VTR sottolineava come l'analisi per essere completa avrebbe dovuto fare riferimento a tre elementi principali: la capacità dell'esercizio di fornire informazioni utili al decisore politico, la sua efficacia nell'allocazione delle risorse e il suo impatto, cognitivo e culturale, sul sistema scientifico nazionale.

In riferimento ai primi due punti, nonostante la VTR fosse stata esplicitamente progettata e realizzata con l'intento di fornire dati adeguati all'allocazione delle risorse, i sui risultati hanno influito in misura marginale su quest'aspetto. In riferimento al finanziamento alle università dal 2006 al 2008 il risultato nella VTR è stato uno dei parametri proposti dal CNVSU per l'assegnazione di una quota del Fondo di Finanziamento Ordinario (FFO), insieme al numero di docenti e ricercatori per struttura e al successo nella presentazione di progetti di ricerca di interesse nazionale (PRIN)[19]. Tuttavia i risultati della VTR hanno inciso sull'assegnazione solo marginalmente, secondo Rebora (2012): per l'1,2% nel 2006, per lo 0,2% nel 2007 e per lo 0,7% nel 2008.

Dal 2009 i risultati della VTR hanno assunto un peso maggiore, nonostante risultassero già abbastanza datati, a seguito della conversione in legge del decreto 180 del 2008 che stabiliva che almeno una quota del 7% del FFO, detta quota premiale, fosse assegnata agli atenei in relazione alla qualità della didattica e della ricerca. Anche qui però dei quattro indicatori riferiti alla ricerca tre sono legati, sostanzialmente, alla capacità di attrazione di risorse, e quello basato sul risultato ottenuto nella VTR è l'unico legato alla *performance*.

La stessa Reale in un contributo recente sottolinea come l'effetto della VTR sui finanziamenti è stato limitato «ed è intervenuto in tempi lontani rispetto a quando la valutazione medesima era stata prodotta, rendendo quindi i parametri poco accettabili» (Reale, 2013, p. 152).

Dal punto di vista della ricezione da parte della comunità scientifica si può dire che la VTR abbia riscontrato un certo interesse[20] ed è stata da più parti ritenuta un successo, le rare critiche rinvenibili risultano centrate sulla selezione e l'operativizzazione degli indicatori piuttosto che sulla valutazione dei prodotti (a titolo di esempio di veda Fabbis e Gnaldi, 2008). E' da notare che CIVR fu formalmente sostituito dall'Anvur nello stesso 2006: la VTR è stata dunque percepita come un esercizio che non si sarebbe ripetuto, perlomeno non con le stesse modalità, e questo ha verosimilmente contribuito a ridurne l'impatto, almeno a livello culturale. Tuttavia la sua realizzazione «ha reso evidente che era non solo concretamente possibile, ma anche opportuno un

[19] La legge 537/1993 stabilisce che il FFO è composto da una quota base, proporzionale alla somma dei trasferimenti statali e delle spese sostenute direttamente dallo Stato negli anni precedenti, e da una quota di riequilibrio, da determinarsi in relazione: ai costi standard di produzione per studente e a obiettivi di qualità della ricerca. La legge stabilisce inoltre che il riparto della quota di riequilibrio debba essere finalizzato a ridurre i differenziali nei costi standard di produzione all'interno delle diverse aree disciplinari e riallineare le risorse fra diverse aree disciplinari. Negli anni sono stati implementati diversi criteri di ripartizione della quota di riequilibrio qui si fa riferimento al meccanismo in vigore tra il 2004 e il 2008.

[20] Secondo Rebora (2012) gli accessi al sito del CIVR in concomitanza con la pubblicazione dei risultati nella VTR sono stati 460.000 nel solo febbraio 2006.

esercizio di valutazione della ricerca nelle università italiane, obiettivo considerato fino ad allora non fattibile e non desiderabile da larga parte del corpo accademico» (Reale, 2013, p. 152).

Il periodo di transizione incluso tra il 2006, anno di pubblicazione dei risultati della VTR e di istituzione dell'Anvur, e il 2010, anno di regolamentazione della nuova agenzia, sembra invece aver prodotto varie forme di resistenza all'idea di una valutazione forte. Il dibattito che ha preceduto e seguito il nuovo esercizio di valutazione della ricerca e il livore (come lo definisce Palumbo, 2013) verso la valutazione in generale, sembrano spiegabili alla luce, da un lato, della sempre maggiore scarsità delle risorse stanziate e delle ripetute dichiarazioni d'intenti, provenienti dai governi quanto dalle parti politiche, circa la necessità di distribuirle secondo un principio di premialità, dall'altro, dal contesto di discontinuità istituzionale e incertezza normativa che per anni ha caratterizzato la valutazione dell'università in Italia.

1.3 L'Agenzia di Valutazione del sistema Universitario e della Ricerca (Anvur) e la valutazione della qualità della ricerca VQR 2004/2010

All'Agenzia di Valutazione del sistema Universitario e della Ricerca (Anvur), fin dalla sua istituzione, venivano attribuiti compiti relativi a diversi ambiti:

«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 Ministero dell'università e della ricerca;

b) indirizzo coordinamento e vigilanza delle attività di valutazione interna 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 innovazione» (legge 286 del 2006, art. 2, comma 138).

L'istituzione dell'Anvur risponde a una serie di obiettivi, soprattutto di natura politica, legati da un lato al processo di Bologna, che richiede la creazione di un'agenzia nazionale di accreditamento e *quality assurance*, dall'altro alla specifica situazione italiana a partire dalla necessità di una razionalizzazione del sistema di valutazione dell'università e della ricerca fino agli adempimenti legati alla valutazione delle performance avviata dalla legge 150/2009 (Reale, 2013).

La struttura e le modalità di funzionamento dell'Agenzia, oltre che la nomina e la durata in carica dei componenti, come previsto dalla stessa legge 286 del 2006 (art. 2, comma 140), dovevano invece essere definite da un regolamento ministeriale, che tuttavia è stato approvato solo dopo quattro anni (DPR n. 76 del 1 Febbraio 2010).

Gli organi dell'Agenzia sono il Presidente, il Consiglio Direttivo e il Collegio dei revisori dei conti, che restano in carica per 4 anni[21]. Il regolamento prevede inoltre la costituzione di un Comitato Consultivo[22], con il compito di fornire pareri al Consiglio Direttivo circa le attività, i criteri e i metodi di valutazione (DPR 76/2010, art. 11), in linea con l'esperienza portata avanti dal CIVR attraverso la creazione dei gruppi di lavoro per la messa a punto delle linee guida della VTR.

Il primo Consiglio Direttivo dell'Anvur, nominato il 22 febbraio 2011[23], riceve nel luglio dello stesso anno il mandato per la realizzazione di un esercizio di valutazione della ricerca (con il DM del 15 luglio 2011) essendo incluso tra i compiti dell'Agenzia quello di proseguire le attività del CIVR anche «innovando rispetto ai metodi e alle procedure da esso adottati» (DPR n. 76 del 1 Febbraio 2010, art. 2, comma 4). Così da quella che doveva essere la Valutazione Quadriennale della Ricerca è nata la Valutazione della Qualità della ricerca, cioè dei risultati della ricerca scientifica effettuata nel periodo 2004-2010 dalle Università Statali e non Statali, dagli Enti di Ricerca pubblici vigilati dal MIUR e da altri soggetti pubblici e privati che svolgono attività di ricerca (questi ultimi su richiesta esplicita e con la partecipazione ai costi dell'esercizio di valutazione).

In base al bando del 7 novembre 2011 l'Anvur «assume a riferimento come modello organizzativo funzionale delle strutture il dipartimento universitario» (Anvur, 2011, p. 1), prevedendo per le strutture non esplicitamente organizzate secondo questo modello che desiderino ottenere una valutazione per le proprie sottostrutture la possibilità di richiederla comunicandone la

[21] Il Consiglio Direttivo è costituito da sette componenti (con almeno due rappresentanti di ciascun genere) nominati con decreto del Presidente della Repubblica su proposta del Ministro, che li seleziona a partire da un elenco (di non meno di dieci e non più di quindici nominativi) definito da un comitato appositamente costruito. Questo 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» (DPR n. 76 del 1 Febbraio 2010, art. 8, comma 3), e valuta anche indicazioni e relativi curricula fornite dagli interessati: istituzioni, accademie, società scientifiche, esperti, istituzioni, organizzazioni degli studenti, parti sociali. Inoltre, diversamente da quanto previsto per i componenti del CIVR, l'incarico presso il Consiglio Direttivo dell'Anvur «è a tempo pieno ed è incompatibile, a pena di decadenza, con qualsiasi rapporto di lavoro, diretto o indiretto, anche a titolo gratuito, instaurato con le istituzioni valutate» (art. 8, comma 5). E' il Consiglio Direttivo a eleggere il Presidente, con il compito di coordinare e assicurare l'unitarietà delle strategie e delle attività dell'Agenzia (art. 7). Il regolamento prevede però che in prima applicazione «previo sorteggio, sono individuati due componenti del Consiglio direttivo che durano in carica tre anni, e tre componenti che durano in carica quattro anni. Gli altri componenti, tra cui il presidente, durano in carica cinque anni» (art. 6, comma 4).

[22] «Il Comitato consultivo è formato da: un componente designato dal Consiglio universitario nazionale; un componente designato dalla Conferenza dei rettori delle università italiane; tre componenti designati dal Consiglio nazionale degli studenti universitari; un componente designato dalla Conferenza dei presidenti degli enti pubblici di ricerca; un componente designato dall'Accademia dei Lincei; quattro rappresentanti delle parti sociali, designati dal Consiglio nazionale dell'economia e del lavoro; un componente designato dalla Conferenza unificata Stato-regioni, città e autonomie locali; un componente straniero ed uno italiano, se presente nel consiglio direttivo dell'ente, designato dall'*European Research Council*; un componente straniero, ed uno italiano, se presente nel consiglio direttivo dell'ente, designato dall'*European University Association*; un componente straniero ed uno italiano, se presente nel consiglio direttivo dell'ente, designato dall'ESIB - the *National Unions of Students in Europe*; un componente designato dal Convegno permanente dei direttori amministrativi e dirigenti delle università italiane; un componente designato dal Segretario generale dell'OCSE» (DPR n. 76 del 1 Febbraio 2010, art. 11, comma 2).

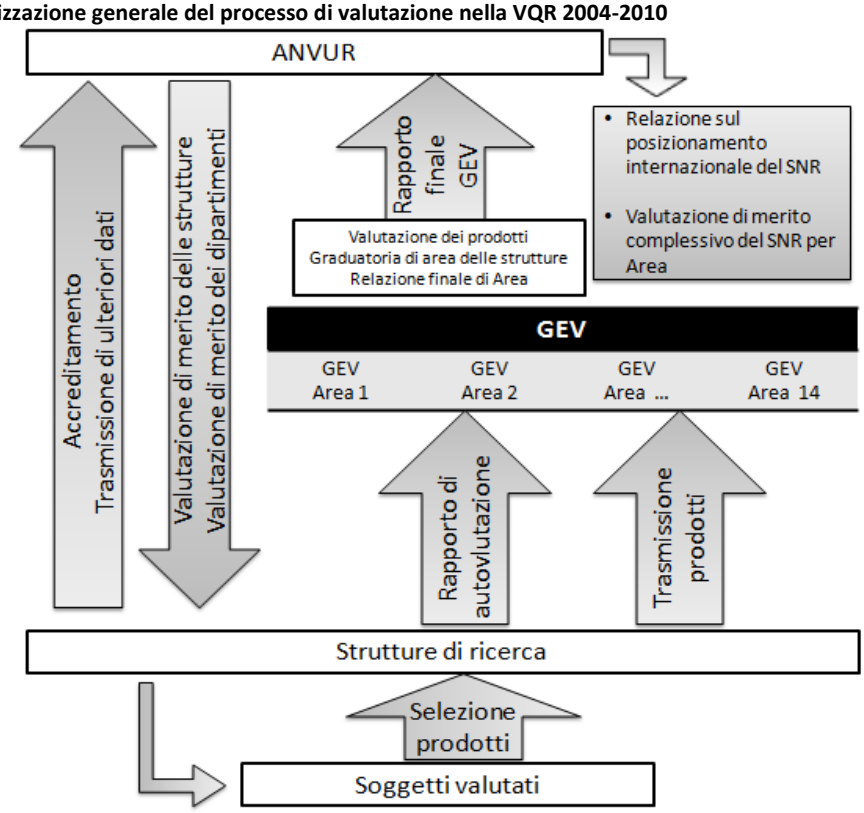
[23] I componenti del primo consiglio direttivo Anvur, in carica nel corso della VQR, erano: Sergio Benedetto, Andrea Bonaccorsi, Massimo Castagnaro, Stefano Fantoni, Giuseppe Novelli, Fiorella Kostoris, Luisa Ribolzi. Viene eletto presidente Stefano Fantoni.

denominazione e la composizione in termini di soggetti valutati. Questi ultimi includono ricercatori (a tempo determinato e indeterminato), assistenti, professori di prima e seconda fascia per le Università, ricercatori e tecnologi per gli Enti di Ricerca.

Come già la VTR anche la VQR si articola sulle 14 Aree disciplinari identificate dal CUN, ma a differenza della prima non definisce nessuna Area speciale. L'Anvur ha costituito un Gruppo di Esperti della Valutazione (GEV) per ciascuna Area, con il compito di valutare i prodotti della ricerca. Le valutazioni sono basate sul metodo della valutazione tra pari e, ove possibile, sulla valutazione diretta tramite analisi bibliometrica (*cfr.* § 1.3.3).

La struttura organizzativa della VQR si differenzia sotto diversi aspetti da quella della VTR (Figura 2), ed è sostanzialmente incentrata su tre livelli di valutazione. Il primo livello in questo esercizio è rappresentato dai soggetti valutati, ricercatori, tecnologi e docenti, che selezionano i prodotti da inviare a valutazione. Il livello intermedio è individuabile nelle strutture[24], che hanno il compito di dirimere le controversie di attribuzione, poiché ciascun prodotto deve essere attribuito a un unico soggetto, pur riferendosi unicamente alle liste fornite dagli autori ed eventualmente (Anvur, 2011). L'ultimo livello è rappresentato dai GEV, che svolgono lo stesso ruolo previsto per i *panel* di area nella VTR, ma ricevono ed elaborano anche i rapporti di autovalutazione delle strutture.

Figura 2 - Organizzazione generale del processo di valutazione nella VQR 2004-2010



[24] Le strutture di ricerca considerate nella VQR sono Atenei ed Enti, con le relative sottostrutture, ad esempio i dipartimenti, che ne costituiscono il modello organizzativo funzionale (Anvur, 2011, p. 1). Stando al bando la struttura: «seleziona i prodotti di ricerca utilizzando unicamente le liste predisposte dai soggetti valutati ad essa afferenti […] avendo cura di risolvere gli eventuali conflitti di attribuzione e attribuendo ogni prodotto ad un solo soggetto valutato» (*ivi*, p. 8). Questo ruolo è svolto in parte dai dipartimenti, ma anche gli Atenei sono intervenuti in questa fase, dato che l'attribuzione doveva essere unica per ciascuna struttura, infatti, potevano verificarsi conflitti di attribuzione tra dipartimenti.

La valutazione anche per questo esercizio non si limita a considerare la qualità dei prodotti della ricerca, ma include una serie di altri indicatori, riferiti ad aspetti come: la capacità delle strutture di attrarre risorse esterne sulla base di bandi competitivi; la mobilità internazionale in entrata e in uscita dei ricercatori; l'alta formazione effettuata dalle strutture; le risorse proprie utilizzate dalla struttura per progetti di ricerca.

Rispetto alla VTR il nuovo esercizio di valutazione tiene conto dell'apertura verso il contesto socio-economico, la così detta terza missione, in sede separata dalla valutazione della ricerca, utilizzando indicatori costruiti *ad hoc*[25]. La valutazione finale delle strutture avviene separatamente per gli indicatori riferiti alla ricerca e alla terza missione, inoltre in riferimento a quest'aspetto il confronto tra strutture intende tenere conto della specificità delle varie aree e delle differenze tra diverse strutture: università generaliste, politecnici, enti di ricerca, eccetera.

Nel bando l'Anvur propone i criteri, gli indicatori e i pesi per la valutazione di strutture e sottostrutture (dipartimenti). I primi fanno riferimento a sette criteri, i secondi solo a quattro, tuttavia l'ordine assegnato ai criteri comuni non varia; in entrambi i casi il peso maggiore viene attribuito al criterio relativo alla qualità della ricerca, seguito dalla capacità di attrazione di risorse, dall'internazionalizzazione e dalla propensione all'alta formazione (Tabella 2 e Tabella 3; *cfr.* Anvur, 2011, Appendice A). I criteri aggiuntivi riferiti alle strutture si riferiscono alla qualità dei prodotti dei soggetti in mobilità, cioè assunti o promossi dalla struttura nel periodo 2004-2010, alla propensione all'utilizzo di risorse proprie per il finanziamento di progetti di ricerca e al miglioramento rispetto al risultato ottenuto nella VTR 2001-2003 (Tabella 2).

Tabella 2 - I criteri di valutazione della VQR per le strutture (adattamento da Anvur, 2011, pp. 14-15)

Qualità della ricerca	Somma delle valutazioni ottenute dai prodotti presentati, espressa come percentuale del valore complessivo dell'Area.	0.5
Capacità di attrazione risorse	Somma dei finanziamenti ottenuti partecipando ai bandi competitivi[26], espressa come percentuale del valore complessivo dell'Area.	0.1
Qualità della produzione dei soggetti in mobilità	Somma delle valutazioni ottenute dai prodotti presentati dai soggetti valutati che, nel periodo 2004-2010, sono stati reclutati dalla struttura o in essa incardinati in una fascia o ruolo superiore.	0.1
Internazionalizzazione	Mobilità (espressa in mesi-persona) dei ricercatori in uscita e in entrata, espressa come percentuale del valore complessivo dell'Area.	0.1
	Somma delle valutazioni ottenute dai prodotti eccellenti con almeno un coautore con afferenza a un ente straniero, espressa come percentuale del valore complessivo dell'Area.	
Propensione all'alta formazione	Numero di studenti di dottorato, assegnisti di ricerca, borsisti post-doc, espresso come percentuale del valore complessivo dell'Area.	0.1
Propensione all'utilizzo di risorse proprie	Somma dei finanziamenti derivati da risorse finanziarie della struttura senza vincoli di destinazione destinate al finanziamento di progetti di ricerca, espressa come percentuale del valore complessivo dell'Area.	0.05
Miglioramento	Differenza della *performance* relativa all'indicatore di qualità della ricerca ottenuta nella VQR 2004-2010 e quella ottenuta dall'analogo indicatore nel VTR 2001-2003.	0.05

[25] La terza missione è concettualizzata come l'apertura verso il contesto socio-economico, esercitata attraverso la valorizzazione e il trasferimento delle conoscenze, e gli indicatori considerati includono: gli importi dei contratti di ricerca o consulenza acquisiti con committenza esterna; i brevetti; gli *spin off* attivati; gli incubatori di impresa compartecipati; i consorzi partecipati; gli scavi archeologici compartecipati; i poli museali; altre attività di terza missione non comprese tra le precedenti.

[26] I bandi competitivi cui l'Anvur fa riferimento sono sia quelli nazionali (PRIN, FIRB, FAR, ASI, PNR) che quelli internazionali (Programmi Quadro dell'Unione Europea, Ente Spaziale Europeo, NIH, ecc.).

Tabella 3 - I criteri di valutazione della VQR per i dipartimenti (adattamento da Anvur, 2011, pp. 17-18)

Qualità della ricerca	Somma delle valutazioni ottenute dai prodotti presentati, espressa come percentuale del valore complessivo dell'Area.	0.5
Capacità di attrazione risorse	Somma dei finanziamenti ottenuti partecipando ai bandi competitivi, espressa come percentuale del valore complessivo dell'Area.	0.2
Internazionalizzazione	Mobilità (espressa in mesi-persona) dei ricercatori in uscita e in entrata, espressa come percentuale del valore complessivo dell'Area.	0.2
	Somma delle valutazioni ottenute dai prodotti eccellenti con almeno un coautore con afferenza a un ente straniero, espressa come percentuale del valore complessivo dell'Area.	
Propensione all'alta formazione	Numero di studenti di dottorato, assegnisti di ricerca, borsisti post-doc, espresso come percentuale del valore complessivo dell'Area.	0.1

L'inclusione dei dipartimenti tra le unità cui riferire la valutazione è una delle principali novità previste per la VQR (Reale, 2013). Nelle intenzioni dell'Anvur la valutazione dei dipartimenti, aggiunta a quella delle strutture già prevista dalla VTR, dovrebbe informare le amministrazioni circa il contributo dei vari dipartimenti (o delle sotto-strutture equivalenti) alla valutazione complessiva della struttura, consentendo di «tenerne conto nel modo che riterranno più appropriato nella distribuzione interna delle risorse» (Anvur, 2011, p. 17). La VQR viene però implementata in una fase di riordino dei dipartimenti che ha fatto seguito alla legge 240/2010, e l'Anvur è costretta a richiedere alle stesse strutture di trasmettere la composizione dei nuovi dipartimenti e dei soggetti che ne fanno parte entro il maggio del 2012 per poter effettuare la valutazione facendo riferimento al nuovo assetto.

Le differenze tra la VQR e l'esercizio precedente sono molteplici, ma la più evidente è sicuramente riferibile alle dimensioni. Innanzitutto l'estensione temporale è più ampia, sette anni anziché tre, e i prodotti sottoposti a valutazione sono notevolmente più numerosi: 184.742 contro i 17.329 della VTR. In termini di risorse umane impiegate, mentre i *panel* di area includevano complessivamente 151 *panelist*, che si sono avvalsi di 6.661 esperti (CIVR, 2006a), i GEV erano composti complessivamente da 436 esperti valutatori e 16 assistenti GEV[27], e il processo di valutazione dei prodotti ha coinvolto 14.770 revisori attivi (di cui 4.620, il 31,3%, con affiliazione straniera) (Anvur, 2013).

Un'altra differenza rilevante è che per questo esercizio le strutture non possono trarre un grande vantaggio dalla possibilità di selezionare liberamente i prodotti: tutti i ricercatori attivi sono tenuti a partecipare all'esercizio di valutazione. In più i soggetti inattivi producono una penalizzazione per la struttura, con una sottrazione di punteggio per ciascun prodotto atteso ma non presentato, dunque: «il numero di pubblicazioni da sottomettere diventa indirettamente un parametro minimo di produttività per i singoli ricercatori» (Reale, 2013, p. 153). Infine nella VQR vengono valutati i ricercatori attivi al momento del bando, non quelli presenti nel periodo valutato, dunque tanto la mobilità quanto i pensionamenti potrebbero incidere sui suoi risultati (Reale, 2013).

Eppure nessuna di queste differenze è stata discussa quanto l'ultima: la VQR per la valutazione dei prodotti della ricerca non utilizza esclusivamente la peer review, ma anche la valutazione diretta tramite analisi bibliometrica. La fase in cui i singoli GEV hanno definito i criteri per

[27] Nel bando del 2011 si prevedeva di utilizzare 450 esperti valutatori e la numerosità di ogni GEV era determinata dall'Anvur in base al numero di prodotti attesi, ma: «di fatto, a causa delle dimissioni di un numero limitatissimo di componenti GEV, e della necessità di integrare la composizione di alcuni GEV, il numero è leggermente superiore» (Anvur, 2013, p. 18).

l'analisi bibliometrica e proposto una classificazione delle riviste è stata forse la più dibattuta dell'intero esercizio (si pensi, fra tutti, alla serie di articoli al vetriolo pubblicati da ROARS[28]). L'uso della bibliometria ha avuto un'incidenza diversa da Area ad Area, e se in qualche caso è stato utilizzato un protocollo relativamente semplice e lineare in altri si è fatto ricorso a soluzioni più complesse (*cfr*. Capitolo 4). La classificazione delle riviste in classi di merito ha invece sollevato un dibattito tale da non essere neppure utilizzata per la valutazione, ma solo per un confronto *ex post*, sostanzialmente informativo, con i risultati della peer review (Anvur, 2013a, Appendice B).

1.3.1 Le aree disciplinari e i gruppi di esperti valutatori (GEV)

I gruppi di esperti valutatori hanno svolto un ruolo centrale nel corso della VQR, non solo a livello esecutivo, ma anche dal punto di vista decisionale.

I membri dei GEV sono stati nominati direttamente dall'Anvur tra coloro che avevano risposto al bando per la segnalazione di disponibilità a partecipare alle procedure di valutazione della VQR 2004-2008 pubblicato dal CIVR nell'estate 2010, estendendo la selezione qualora non vi fosse un numero sufficiente di candidati rispondenti ai criteri fissati (Anvur, 2013).

La selezione iniziale si è basata su tre criteri fondamentali:

«1. qualità scientifica (tenendo conto del merito scientifico, delle sedi di pubblicazione, del numero delle citazioni, dell'impatto della ricerca nella comunità internazionale e di eventuali premi di ricerca o altri riconoscimenti);
2. continuità della produzione scientifica negli ultimi 5 anni;
3. esperienza in attività di valutazione a livello nazionale e internazionale» (Anvur, 2013, p. 18).

Coloro che rispondevano a questi tre criteri venivano ulteriormente selezionati, con lo scopo di rispettare una serie di condizioni per ciascun GEV:

«a. copertura delle linee culturali e di ricerca all'interno delle aree;
b. percentuale significativa di docenti di università straniere, con l'obiettivo globale medio di 20% di docenti con queste caratteristiche;
c. attenzione alla distribuzione di genere;
d. equa distribuzione di sede, ove possibile, per i candidati di atenei e enti di ricerca italiani;
e. equa distribuzione geografica, ove possibile, per i candidati di atenei e enti di ricerca italiani» (Anvur, 2013, p. 19).

Il numero di componenti di ciascun GEV era stabilito dall'Anvur nel bando del novembre 2011 sulla base del numero di prodotti attesi per l'Area CUN di riferimento (Tabella 4). I GEV avevano la facoltà di nominare, se opportuno e con l'approvazione dell'Anvur, dei sotto-gruppi, detti Sub-GEV, più omogenei dal punto di vista disciplinare, e in diversi casi le valutazioni sono state presentate

[28] *Return On Academic ReSearch* www.roars.it, si tratta di un sito web creato da un gruppo di accademici, ricercatori, esperti di valutazione, studenti, con lo scopo di portare contributi alla discussione circa i processi che hanno investito negli ultimi anni l'università italiana.

facendo riferimento a questi sottoinsiemi se non direttamente ai settori scientifico-disciplinari (SSD)[29].

Tabella 4- Numerosità del GEV per Area (Anvur, 2011, pp. 2 e 3).

Area		Numerosità GEV
Area 1	Scienze matematiche e informatiche	25
Area 2	Scienze fisiche	18
Area 3	Scienze chimiche	23
Area 4	Scienze della terra	9
Area 5	Scienze biologiche	38
Area 6	Scienze mediche	79
Area 7	Scienze agrarie e veterinarie	24
Area 8	Ingegneria civile e architettura	28
Area 9	Ingegneria industriale e dell'informazione	40
Area 10	Scienze dell'antichità, filologico-letterarie e storico-artistiche	42
Area 11	Scienze storiche, filosofiche, psicologiche e pedagogiche	38
Area 12	Scienze giuridiche	37
Area 13	Scienze economiche e statistiche	36
Area 14	Scienze politiche e sociali	13

Ciascun GEV aveva il compito di definire, di concerto con l'Anvur i criteri e le procedure da utilizzare per la valutazione dei prodotti della ricerca. Il giudizio di qualità su ciascun prodotto, che include una parte descrittiva, poteva essere formulato utilizzando diversi approcci:

a) la valutazione diretta da parte del GEV, da condurre anche utilizzando l'analisi bibliometrica, basata sulle citazioni del prodotto e sul fattore di impatto della rivista ospitante il prodotto (ove questa risultasse applicabile);
b) la *peer review* affidata a esperti esterni, indipendenti, scelti dal GEV con il compito di esprimersi, in modo anonimo, sulla qualità delle pubblicazioni selezionate.

Tra le prerogative dei singoli GEV vi era quella di decidere la percentuale di prodotti cui applicare l'analisi bibliometrica, anche se il bando stabiliva dei parametri: «almeno la metà più uno dei prodotti complessivi (includendo tutte le aree) sarà valutata utilizzando la peer review; nel caso di prodotti la cui valutazione sia affidata alla peer review, i GEV si atterranno al criterio generale di distribuire tali prodotti sul massimo numero di soggetti valutati» (Anvur, 2011, p. 7).

Il margine di autonomia dei GEV era abbastanza ampio anche in riferimento all'interpretazione e alla modulazione dei criteri già definiti dal DM del luglio 2011 e dal bando; il *Rapporto finale Anvur* sottolinea gli elementi comuni ai vari GEV, mettendo in luce anche gli elementi specifici per i quali ogni GEV ha scelto la via più rispondente alle peculiarità delle discipline in esso rappresentate[30] (Anvur, 2013a, pp. 21-23).

Il primo elemento comune a tutti i GEV è l'attribuzione della responsabilità finale della valutazione dei prodotti con l'assegnazione dei prodotti alle classi di merito al GEV stesso. Tutti i GEV hanno scelto di utilizzare a questo fine l'*informed* peer review, che permette di tener conto di più elementi di valutazione per la classificazione finale di merito, anche se questi elementi variano da GEV a GEV in relazione alle caratteristiche delle discipline di riferimento.

[29] Gli SSD sono i 370 Settori Scientifico-Disciplinari nei quali si articolano le quattordici Aree CUN, definiti con il DM del 4 ottobre 2000 e aggiornati con il DM del 18 marzo 2005.

[30] I criteri dei singoli GEV e un documento di accompagnamento redatto da Anvur sono stati pubblicati il 29 febbraio 2012 (Anvur, 2012).

Sulla peer review era basata la valutazione di tutti prodotti della ricerca, a eccezione degli articoli su rivista, e tanto per le monografie quanto per i contributi in volume era previsto l'uso della *informed* peer review*.* Per queste procedure era prevista la predisposizione di una scheda di revisione che prevedeva tre domande a risposta multipla, una per ciascun criterio di qualità, pesate a seconda della rilevanza attribuita al criterio di riferimento. Comuni a tutti i GEV erano anche la procedura per l'individuazione dei revisori esterni e le norme atte a evitare i conflitti di interesse.

I GEV che potevano avvalersi di basi di dati adeguate per la valutazione bibliometrica dei prodotti, in particolare di Web of Science (WoS) e Scopus[31], hanno tutti previsto l'utilizzo di due indicatori, uno legato alla rivista in cui il prodotto era collocato, l'altro alle citazioni ricevute, e optato per l'*informed* peer review nei casi in cui questi due indicatori divergessero. Il GEV di Scienze statistiche ed economia (13) ha preferito l'applicazione di un algoritmo di valutazione differente, con un diverso peso tra indicatore bibliometrico e indicatore citazionale, ma tutti i GEV che hanno utilizzato la bibliometria hanno adattato l'algoritmo di valutazione alle proprie esigenze specifiche (*cfr*. Capitolo 4).

L'elemento comune ai GEV legati a discipline che non dispongono di banche dati sufficientemente affidabili[32] è stato l'utilizzo generalizzato della peer review per la valutazione di tutti i prodotti di ricerca. Inoltre sono state elaborate delle classificazioni delle riviste per gli SSD di competenza del GEV, che tuttavia non sono state utilizzate ai fini della valutazione.

1.3.2 I prodotti sottoposti a valutazione

L'esercizio di valutazione, come già rilevato, considera un numero elevatissimo di prodotti della ricerca, che includono:

a) articoli su riviste;
b) libri, capitoli di libri e atti di congressi (solo se dotati di ISBN);
c) edizioni critiche, traduzioni e commenti scientifici;
d) brevetti concessi nel settennio di cui risulti autore/coautore il soggetto valutato;
e) composizioni, disegni, *design*, *performance*, mostre ed esposizioni organizzate, manufatti, prototipi e opere d'arte e loro progetti, banche dati e *software*, carte tematiche, se corredati da pubblicazioni atte a consentirne adeguata valutazione (Anvur, 2011).

I prodotti da sottoporre a valutazione vengono selezionati dalla struttura di appartenenza dell'autore, a partire da un elenco stilato dall'autore stesso che riporti le pubblicazioni in ordine di priorità. I prodotti con coautori appartenenti a diverse strutture potevano essere presentati da ciascuna delle strutture di afferenza degli autori. Contrariamente a quanto avveniva nella VTR il peso dei prodotti frutto della collaborazione fra strutture diverse non viene ridotto tramite un coefficiente

[31] I GEV di Scienze matematiche e informatiche (1), Scienze fisiche (2), Scienze chimiche (3), Scienze della terra (4), Scienze biologiche (5), Scienze mediche (6), Scienze agrarie e veterinarie (7), Ingegneria industriale e dell'informazione (9), più parti dei GEV di Ingegneria civile e architettura (8) e Scienze storiche, filosofiche, psicologiche e pedagogiche (11).

[32] Parte dei GEV di Ingegneria civile e architettura (8) e Scienze storiche, filosofiche, psicologiche e pedagogiche (11), più i GEV di Scienze dell'antichità, filologico-letterarie e storico-artistiche (10), Scienze giuridiche (12), Scienze economiche e statistiche (13), Scienze politiche e sociali (14).

di proprietà e conta per ciascuna delle strutture che lo presentano quanto tutti gli altri prodotti.

Gli autori, indicati nella terminologia della VQR come i *soggetti valutati* (Anvur, 2011, p. 3), sono ricercatori, professori associati e professori ordinari delle Università, ricercatori, primi ricercatori, dirigenti di ricerca e tecnologi, primi tecnologi e dirigenti tecnologi degli Enti di Ricerca, in servizio alla data del bando (novembre 2011). I soggetti valutati sono considerati appartenenti alla struttura[33] presso cui lavorano alla data del bando, indipendentemente dalle affiliazioni precedenti, e i loro prodotti sono attribuiti a quella stessa struttura, indipendentemente dall'affiliazione del suo autore al momento della pubblicazione.

Il numero di prodotti da sottomettere a valutazione viene definito dall'Anvur nel bando VQR, ed è legato al tipo di struttura cui l'autore appartiene, al suo ruolo all'interno della struttura e al tempo trascorso dalla sua entrata in servizio (Tabella 5)[34].

Tabella 5 – Prodotti attesi per caratteristiche dei soggetti valutati (Anvur, 2011, p. 5)

Ruolo	Restrizione	N. prodotti	Struttura
Professore ordinario		3	Università
Professore associato		3	Università
Ricercatore universitario	se in servizio da prima del 1/1/2006	3	Università
	se è tra 1/1/2006 e 31/12/2007	2	Università
	se è tra 1/1/2008 e 31/12/2009	1	Università
	se è successiva al 1/1/2010	0	Università
Dirigente di Ricerca		6	Ente di ricerca
Primo ricercatore		6	Ente di ricerca
Ricercatore presso Ente di ricerca	se in servizio da prima del 1/1/2006	6	Ente di ricerca
	se è tra 1/1/2006 e 31/12/2007	4	Ente di ricerca
	se è tra 1/1/2008 e 31/12/2009	2	Ente di ricerca
	se è successiva al 1/1/2010	0	Ente di ricerca
Dirigente tecnologo		3	Ente di ricerca
Primo tecnologo		3	Ente di ricerca
Tecnologo	se in servizio da prima del 1/1/2006	3	Ente di ricerca
	se è tra 1/1/2006 e 31/12/2007	2	Ente di ricerca
	se è tra 1/1/2008 e 31/12/2009	1	Ente di ricerca
	se è successiva al 1/1/2010	0	Ente di ricerca
Professore Ordinario incaricato di ricerca presso Ente di ricerca per almeno 3 anni		3	Ente di ricerca
Professore Associato incaricato di ricerca presso Ente di ricerca per almeno 3 anni		3	Ente di ricerca
Ricercatore Universitario incaricato di ricerca presso Ente di ricerca per almeno 3 anni		3	Ente di ricerca

Nonostante quasi tutti i soggetti valutati fossero tenuti a sottoporre alla valutazione parte della propria produzione scientifica, la valutazione non può essere riferita ai singoli autori. Infatti la scelta dei prodotti da associare a ciascun soggetto valutato, effettuata in caso di conflitti di attribuzione dalle strutture, è ragionevolmente mirata all'ottimizzazione del risultato della struttura e non del singolo soggetto. Inoltre in diversi settori disciplinari il numero di prodotti richiesti risulta

[33] Al fine di non appesantire il testo di qui in avanti il termine struttura è riferito sia agli Atenei e agli Enti, sia alle loro sottostrutture, salvo precisazioni.

[34] Nel caso di periodi di congedo avvenuti nel periodo di riferimento della VQR il numero di prodotti da presentare veniva ridotto di 1/3 rispetto al valore previsto per congedi di durata complessiva compresa fra 2 e 4 anni, e di 2/3 per congedi di durata complessiva compresa fra 4 e 6 anni. Infine per congedi superiori ai 6 anni i soggetti erano esentati dall'esercizio di valutazione.

molto parziale rispetto alla produzione complessiva di un autore per un settennio, e la procedura di valutazione non tiene conto della dimensione del gruppo di ricerca.

I prodotti dovevano essere trasmessi ai GEV per via telematica[35], accompagnati da una scheda descrittiva. Quest'ultima doveva includere una serie di informazioni:

«1. Metadati bibliografici del prodotto;
2. Identificazione del soggetto valutato di riferimento;
3. Identificazione dell'area e del settore scientifico disciplinare;
4. Indicazione della presenza di almeno un coautore con afferenza ad un ente straniero;
5. Abstract del prodotto;
6. La eventuale segnalazione, a cura della struttura, che il prodotto proviene da attività di ricerca in aree emergenti a livello internazionale o in aree di forte specializzazione o a carattere interdisciplinare, per le quali si suggerisce l'adozione preferenziale della metodologia di peer review in ragione della minore rappresentazione di tali aree negli indicatori bibliometrici;
7. Ogni altra informazione che si ritenga utile alla valorizzazione del prodotto (premi ricevuti, autorevolezza della rivista/editore, ecc.)» (Anvur, 2011, p.6).

L'associazione del prodotto all'Area e al SSD era dunque decisa dalle strutture, anche indipendentemente da quelle di riferimento del soggetto valutato cui era attribuito.

Complessivamente risulta mancante, rispetto al numero atteso, il 5.1% dei prodotti, ma la variabilità di questa quota è piuttosto elevata, non solo in ragione di una diversa incidenza dei soggetti inattivi, ma anche della assegnazione alle Aree disciplinari operata dalle strutture (Tabella 6).

Tabella 6 - Prodotti attesi e conferiti per Area del soggetto valutato e Area del prodotto (Anvur, 2013a, Tab.2.4-6)

Area	Prodotti attesi (Area del soggetto valutato)	Prodotti conferiti (Area del soggetto valutato)		Prodotti conferiti (Area del prodotto)		Saldo Prodotti conferiti
		N.	*% mancanti*	N.	% mancanti	
01	11.752	10.685	9,1	9.682	17,6	-1.003
02	20.286	19.773	2,5	19.386	4,4	-387
03	11.933	11.608	2,7	11.812	1,0	204
04	8.859	8.433	4,8	7.229	18,4	-1.204
05	17.268	16.407	5,0	17.298	-0,2	891
06	29.454	26.713	9,3	27.085	8,0	372
07	10.349	10.004	3,3	9.866	4,7	-138
08	9.934	9.533	4,0	9.657	2,8	124
09	16.858	16.347	3,0	17.654	-4,7	1.307
10	14.637	14.073	3,9	13.966	4,6	-107
11	13.487	13.152	2,5	13.158	2,4	6
12	12.798	11.882	7,2	11.886	7,1	4
13	12.654	11.941	5,6	11.765	7,0	-176
14	4.494	4.327	3,7	4.434	1,3	107
Totale	**194.763**	**184.878**	**5,1**	**184.878**	**5,1**	**0**

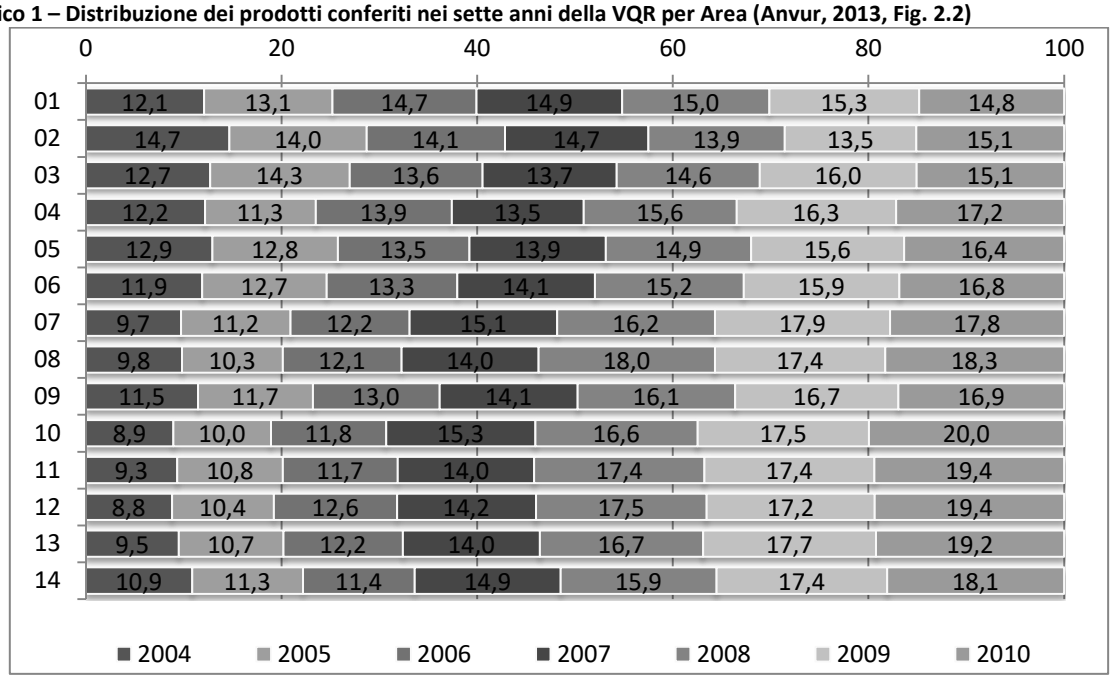
In Tabella 6, ad esempio, per l'Area 1 osservando l'Area del soggetto valutato è mancante il 9,1% dei prodotti, ma questa percentuale quasi raddoppia in riferimento all'Area del prodotto. L'Area

[35] Questa come tutte le altre procedure telematiche della VQR, e come quelle della VTR, è stata messa a punto e gestita dal consorzio interuniversitario CINECA. In casi particolari di indisponibilità del prodotto in formato elettronico, era naturalmente ammesso l'invio in formato cartaceo, a seguito della comunicazione ai GEV e del relativo consenso.

5 e l'Area 9 presentano una percentuale di mancanti rispetto all'Area del prodotto negativa, ricevono cioè una quantità di prodotti di soggetti appartenenti ad altre Aree tale da bilanciare e superare il numero di prodotti mancanti dei propri soggetti. Il saldo tra prodotti conferiti *da* altre Aree e prodotti conferiti *ad* altre Aree risulta fortemente negativo per Scienze matematiche e informatiche (1) e Scienze della terra (4), fortemente positivo per Scienze biologiche (5) e Ingegneria industriale e della informazione (9). Le caratteristiche delle Aree disciplinari coinvolte lasciano ipotizzare che i prodotti usciti dall'Area 1 finiscano nell'Area 9, mentre quelli usciti dall'Area 4 finiscano nell'Area 5, e i dati sembrano confermare in una certa misura queste dinamiche (Anvur, 2013a, Tab. 2.8 e 2.9).

I prodotti conferiti per la valutazione sono più numerosi per gli ultimi anni del periodo di riferimento dell'esercizio (Grafico 1), ma la distribuzione varia fortemente da Area ad Area. A fronte di una sostanziale equidistribuzione lungo l'asse temporale per le Aree di Scienze fisiche (2), Scienze chimiche (3) e Scienze della terra (4), altre Aree presentano una quota doppia di prodotti pubblicati nel 2010 rispetto a quella dei prodotti del 2004, ad esempio come Scienze agrarie e veterinarie (7), Ingegneria civile e architettura (8), Scienze dell'antichità, filologico-letterarie e storico-artistiche (10), Scienze storiche, filosofiche, psicologiche e pedagogiche (11), Scienze giuridiche (12) e Scienze economiche e statistiche (13).

Grafico 1 – Distribuzione dei prodotti conferiti nei sette anni della VQR per Area (Anvur, 2013, Fig. 2.2)



Potrebbe apparire singolare che nelle Aree in cui la produzione scientifica è a impatto breve, cioè nelle scienze dure come fisica e chimica, si siano sottoposti a valutazione prodotti equidistribuiti nel tempo, mentre in Aree in cui la produzione ha un impatto molto più lungo e in alcuni casi potenzialmente indefinito, come per letteratura e filologia, oppure per le scienze storiche e filosofiche, i soggetti valutati abbiano preferito presentare una produzione più recente. Al contrario è possibile individuare diverse possibili spiegazioni.

In primo luogo nelle Aree in cui era previsto l'uso dell'analisi bibliometrica è immaginabile che i soggetti valutati abbiano scelto i prodotti da inviare a valutazione sulla base dei valori degli indicatori utilizzati dal GEV di riferimento e che la loro scelta dipenda dal riscontro, per le

pubblicazioni meno recenti, di valori più elevati sugli indicatori, in conseguenza della maggiore maturità dei relativi dati citazionali[36]. Nelle Aree non bibliometriche, invece, in assenza di fonti di dati i soggetti potrebbero aver selezionato più liberamente i propri prodotti, ed è plausibile che abbiano ritenuto quelli più recenti di maggiore qualità perché riferibili a fasi più avanzate del proprio lavoro.

Né un effetto della VTR né un effetto annuncio relativo alla stessa VQR sembrano plausibili, poiché date le caratteristiche del primo esercizio la possibilità di un effetto sulla produzione scientifica sembra estremamente remota, mentre i criteri per la VQR sono stati pubblicati solo nel 2011. Neppure l'Abilitazione Scientifica Nazionale(ASN)[37] dovrebbe aver influito sulla produzione scientifica o sulla selezione dei prodotti, dato che la sua istituzione è del 2010 e i relativi criteri sono stati pubblicati solo nel 2012.

1.3.3 La valutazione della qualità dei prodotti

Gli aspetti che definiscono il concetto di qualità della ricerca sono così espressi nel decreto ministeriale e nel bando di partecipazione alla VQR del 2011 (DM 17 del 15 luglio 2011, art. 8, comma 2; Anvur, 2011, p. 7):

a) *rilevanza*: «da intendersi come 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»;
b) *originalità/innovazione*: «da intendersi come contributo all'avanzamento di conoscenze o a nuove acquisizioni nel settore di riferimento»;
c) *internazionalizzazione*: «da intendersi come posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica, inclusa la collaborazione esplicita con ricercatori e gruppi di ricerca di altre nazioni»[38].

L'obiettivo finale della valutazione è l'assegnazione di ciascun prodotto a una *classe di merito*. Questa classificazione fa riferimento alla collocazione dei prodotti nella «scala di valore condivisa a livello internazionale» (Anvur, 2011a, p. 7). Stando al bando:

«1. I prodotti di livello eccellente sono quelli riconosciute 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.
2. I prodotti di livello buono sono quelli di importanza internazionale e nazionale riconosciute per originalità dei risultati e rigore metodologico.

[36] E' noto infatti che, nonostante una certa variabilità tra i campi disciplinari, l'affidabilità del numero di citazioni ricevute come indicatore di qualità è bassissima a ridosso della pubblicazione e dopo aver raggiunto un massimo (in un tempo che varia a seconda della disciplina), decresce progressivamente (di nuovo, con una velocità diversa seconda della disciplina) (per tutti Garfield, 1979b; *cfr*. § 4.2.3 e § 6.2.1).

[37] L' ASN, ai sensi dell'articolo 16 della legge 30 dicembre 2010, n. 240, costituisce un requisito necessario per l'accesso al ruolo di prima e seconda fascia per il reclutamento di personale da parte delle Università. Si tratta di una valutazione, svolta da commissioni nazionali, che attesta la qualificazione scientifica dei candidati. Tra i criteri, per le Aree non bibliometriche, vi è il numero di articoli pubblicati in riviste di fascia A (DM 76/2012).

[38] In aggiunta a questi criteri, per i brevetti i giudizi dovevano contenere anche riferimenti al «*trasferimento, allo sviluppo tecnologico e alle ricadute socio-economiche (anche potenziali)»* (Anvur, 2011, p. 7).

3. 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.
4. 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» (*ibidem*).

Ogni prodotto era assegnato a uno o due componenti del GEV, a seconda dell'Area, cui veniva attribuita la responsabilità della valutazione della sua qualità. In base all'Area le procedure adottate sono differenti oltre che l'approccio, bibliometrico o meno, utilizzato per la valutazione, tuttavia il giudizio sintetico finale è sempre espresso dal GEV e concorre a determinare la valutazione della struttura nello stesso modo. A ciascuna classe di merito corrisponde, oltre che un *range* specifico della scala di valore, un punteggio incluso tra 0 e +1:

a) *Eccellente*: nel 20% superiore della scala (peso 1);
b) *Buono*: nel segmento 60% - 80% (peso 0.8);
c) *Accettabile*: nel segmento 50% - 60% (peso 0.5);
d) *Limitato*: nel 50% inferiore (peso 0; Anvur, 2011, p. 7-8).

Nel caso in cui il prodotto appartenesse a tipologie escluse dall'esercizio o presentasse allegati e/o documentazione inadeguati per la valutazione, o ancora fosse pubblicata in anni precedenti o successivi al settennio di riferimento era classificato come *Non valutabile*, con una conseguente assegnazione di un punteggio di penalizzazione (peso -1). E' inoltre prevista una penalizzazione più forte nei casi accertati di plagio o frode, cui viene attribuito un punteggio pari a -2 (*ibidem*).

Un caso particolare è quello dei prodotti con più autori, dato che ciascun prodotto doveva essere assegnato univocamente dalla struttura a un solo soggetto valutato, ma era possibile che diversi soggetti valutati venissero collegati a uno stesso prodotto. Nel caso in cui due strutture dello stesso tipo (ad esempio due diversi Atenei) avessero presentato lo stesso prodotto, ciascuna riferendolo a uno dei propri soggetti valutati, non sarebbe stata applicata nessuna penalizzazione, al fine di: «incentivare, per il futuro, la collaborazione tra strutture diverse» (*ibidem*, p. 6). Nel caso una stessa struttura (Ateneo o Ente) avesse presentato più volte lo stesso prodotto questo sarebbe stato valutato una sola volta, e a ciascuno dei suoi doppi sarebbe stata assegnata una penalizzazione pari a -0.5 (Anvur, 2013a)[39].

Era inoltre possibile che due strutture di diverso tipo (ad esempio un Ateneo e un Ente) associassero lo stesso prodotto allo stesso soggetto valutato[40]; in questo caso il prodotto veniva valutato e al suo doppio era assegnata una penalizzazione (pari a -0.5), dividendo però il punteggio ottenuto tra le due strutture.

Di seguito vengono brevemente descritte le procedure di valutazione così come sono presentate nel Report finale Anvur, l'intento è esporre sinteticamente le loro caratteristiche principali, mentre l'analisi metodologica vera e propria è presentata nei Capitoli 3, 4 e 5.

[39] Nel bando, mentre nel caso di due strutture di due tipi diversi o di un'unica struttura si sottolinea che il prodotto viene valutato una sola volta non vi è una indicazione precisa nel caso le strutture siano entrambi Atenei (Anvur, 2011). E' il caso inoltre di sottolineare che l'algoritmo utilizzato è stato: «deciso congiuntamente da tutti i GEV per attribuire le penalizzazioni» e non ricalca lo schema del bando, ma, che interpreta in maniera favorevole alle strutture il dettato del DM e del bando» (Anvur, 2013a, p. 29).

[40] Nel caso in cui il soggetto valutato afferisca a una università ma abbia un formale incarico di ricerca presso un Ente (ancora attivo alla data del bando; Anvur, 2011).

1.3.3.1 La valutazione tramite peer review

La procedura di valutazione tramite peer review ha previsto la costituzione di un albo di revisori Anvur suddiviso per GEV, poiché nell'albo dei revisori CINECA utilizzato dal MIUR per la valutazione ex ante di PRIN e FIRB:

(a) le credenziali scientifiche dei revisori non erano state sottoposte ad alcuna valutazione;

(b) non vi era un numero sufficiente di revisori stranieri.

La selezione dei revisori è avvenuta, prima sulla base dell'Albo CINECA poi interpellando individualmente gli esperti, sulla base di criteri di merito scientifico tra cui per le Aree cosiddette bibliometriche l'indice *h* di Hirsch e il numero di citazioni, per le altre la produzione scientifica recente (*cfr.* § 5.1.1). Il Consiglio Direttivo dell'Anvur aveva inoltre pubblicato un modulo di auto-candidatura per soggetti che pur non essendo già presenti nell'Albo CINECA intendessero contribuire al processo di valutazione, naturalmente anche queste candidature venivano selezionate dai GEV di riferimento. L'albo ottenuto tramite queste procedure, costituito da oltre 16.000 nomi, è stato ulteriormente integrato durante la fase di valutazione al fine di ottenere la massima copertura delle competenze necessarie alla revisione dei prodotti conferiti.

In relazione a ciascun prodotto, sotto la propria responsabilità, i membri GEV erano tenuti a scegliere separatamente (almeno) due revisori. I documenti relativi ai criteri di valutazione ponevano una notevole enfasi sulla necessità, in questa fase, di evitare i conflitti d'interesse[41]. Diversi prodotti sono stati revisionati dagli stessi membri dei GEV, in particolare quelli classificati in maniera discordante dagli indici bibliometrici.

I revisori avevano a disposizione per la valutazione dei prodotti, una scheda che conteneva una domanda a risposta multipla per ciascuno dei criteri indicati nel bando del 2011: *rilevanza*, *originalità/innovazione, internazionalizzazione* (Anvur, 2011, p. 7; *cfr.* § 4.1.1 e 4.1.2).

I giudizi descrittivi erano funzionali alla formulazione di un giudizio sintetico, mirato all'attribuzione di un livello di merito a ciascun prodotto, dunque la somma dei punteggi ottenuti dal prodotto sulla base delle risposte dei revisori generava una classificazione finale in quattro classi tramite il confronto con dei punteggi-soglia. In riferimento ai contenuti delle schede l'Anvur rimanda ai report delle singole Aree (Anvur, 2013d), tuttavia solo nel caso dell'Area 13, è disponibile sia la scheda sia l'assegnazione delle classi di merito per i punteggi (è riportato in Appendice E, p. 112-113), per l'Area 7 (in Appendice A p. 28) è possibile esaminare la scheda, ma non la corrispondenza dei punteggi con le classi di merito, per tutte le altre Aree la scheda non è stata pubblicata.

Una volta compilata la scheda, la classe di merito corrispondente alle risposte fornite veniva proposta al revisore «per consentirgli di confrontarla con la definizione delle classi […] e, eventualmente, di modificare i punteggi» (Anvur, 2013a, p. 26). La classe di merito finale per ciascun prodotto veniva assegnata dai GEV sulla base dei punteggi espressi dai revisori utilizzando un procedimento predefinito (*cfr.* § 4.3.2). Era prevista la possibilità di richiedere una terza revisione per i casi in cui le valutazioni di due peer risultassero discordanti di più di una classe.

[41] A titolo di esempio si vedano: p. 15 dei Criteri del GEV 1, p. 7 dei Criteri del GEV 3, p. 11 dei Criteri del GEV 11, p. 6 dei Criteri dell'Area 14. Ogni GEV ha proposto delle procedure per la gestione de conflitti di interesse, tuttavia la loro definizione risulta sostanzialmente comune a tutte le quattordici Aree; tutti i documenti relativi ai Criteri sono consultabili sul sito istituzionale dell'Anvur alla pagina: http://www.Anvur.org/index.php?option=com_content&view=article&id=32&Itemid=372&lang=it .

I GEV di Scienze dell'antichità, filologico-letterarie e storico-artistiche (10), Scienze giuridiche (12) e Scienze politiche e sociali (14) hanno utilizzato esclusivamente la revisione dei pari per la valutazione dei prodotti (se la percentuale è inferiore al 100% è a causa dei prodotti classificati come non valutabili), mentre per Scienze fisiche (2) e Scienze chimiche (3) la quota di prodotti sottoposti alla revisione dei pari è inferiore al 30% (da considerare che per le aree bibliometriche circa il 10% dei prodotti è stato valutato con entrambi gli approcci per consentire un confronto dei loro risultati) (Tabella 7).

Tabella 7 – Prodotti attesi, conferiti, e sottomessi alla peer review (Anvur, 2013a, Tab. 3.3)

Area	Prodotti attesi	Prodotti conferiti	Prodotti sottomessi alla peer review	
			N.	% sui prodotti conferiti
01	11.752	10.685	5.180	48,5
02	20.286	19.773	5.350	27,1
03	11.933	11.608	2.879	24,8
04	8.859	8.433	3.390	40,2
05	17.268	16.407	4.985	30,4
06	29.454	26.713	10.330	38,7
07	10.349	10.004	4.501	45,0
08	9.934	9.533	7.541	79,1
09	16.858	16.347	7.351	45,0
10	14.637	14.073	13.942	99,1
11	13.487	13.152	11.186	85,1
12	12.798	11.882	11.784	99,2
13	12.654	11.941	6.277	52,6
14	4.494	4.327	4.304	99,5
Totale	**194.763**	**184.878**	**99.000**	**53,5**

1.3.3.2 La valutazione bibliometrica

La valutazione bibliometrica ha riguardato esclusivamente gli articoli pubblicati su riviste indicizzate nelle basi di dati WoS, di Thomson Reuters, e Scopus, di Elsevier. Le informazioni bibliometriche sono state acquisite dall'Anvur per gli anni dal 2004 al 2011 in riferimento alla produzione scientifica mondiale. La scelta di utilizzare entrambe le basi dati, non esente da complicazioni, è stata portata avanti dall'Agenzia sia con l'intento di evitare di legarsi a un solo gestore, sia perché la copertura delle due basi dati è differente, dunque possono essere considerate parzialmente complementari.

Tutti i GEV legati a discipline valutabili con questo approccio hanno utilizzato un algoritmo simile, che prevedeva una classificazione basata su due indicatori per ogni prodotto:

1) il numero di citazioni;
2) l'*impact factor* per WoS, l'indice SJR (*SCImago Journal Ranking*) di Scopus o altri indicatori circa l'impatto delle riviste (*cfr*. § 4.1.3).

La procedura prevedeva poi il calcolo delle distribuzioni cumulative dei due indicatori all'interno di categorie disciplinari omogenee[42] per anno di pubblicazione, utilizzando le due basi dati WoS e Scopus complete (cioè non limitate ai *record* nazionali). Queste distribuzioni dovevano essere suddivise in quattro classi, ciascuna contenente una percentuale data di riviste o articoli (la classe più

[42] Le *Subject Category* di WoS e le classi dell'ASJC (*All Science Journal Classification*) in Scopus.

elevata il 20% più in alto nella distribuzione, la successiva dal 60% - 80%, la penultima dal 50% - 60%, e l'ultima il rimanente 50%). A ciascun prodotto veniva assegnata, per ciascun indicatore, la classe di merito corrispondente alla sua posizione nella distribuzione cumulata (*cfr*. § 4.2.3).

La sintesi dei due indicatori era legata a una matrice 4x4 data dall'incrocio delle due classificazioni; a ciascuna delle 16 celle corrispondeva una classe di merito, oppure, nel caso di forte divergenza tra gli indicatori era prevista una classe *undecided*. I singoli GEV hanno stabilito autonomamente le regole specifiche per attribuire al prodotto una classe di merito, oppure classificarlo come *undecided* (*cfr*. § 4.3.3). Tuttavia in tutti i GEV i prodotti classificati come *undecided* venivano sottoposti a una seconda valutazione tramite peer review.

L'area di Scienze chimiche (3) con il 97,4% e quella di Scienze biologiche (5) con il 92,4% presentano le quote più elevate di prodotti sottoposti ad analisi bibliometrica (Tabella 8). Vale la pena di sottolineare che la tabella riportata attribuisce i prodotti alle Aree di afferenza dei soggetti valutati, di conseguenza alcuni prodotti delle Aree 10, 12 e 14 valutati tramite analisi bibliometrica, sono quelli assegnati dalle strutture a un GEV diverso da quello degli autori (Anvur, 2013a).

Tabella 8 - Prodotti attesi, conferiti, e sottomessi alla peer review (Anvur, 2013a, Tab. 3.5)

Area	Prodotti attesi	Prodotti conferiti	Valutati tramite bibliometria		Valutati *undecided*	
			N.	% sui conferiti	N.	% su valutati tramite bibliometria
01	11.752	10.685	8.290	77,6	2.198	26,5
02	20.286	19.773	18.164		2.298	12,7
03	11.933	11.608	11.302	97,4	1.574	13,9
04	8.859	8.433	6.544	77,6	1.127	17,2
05	17.268	16.407	15.166	92,4	2.796	18,4
06	29.454	26.713	22.717	85,0	4.328	19,1
07	10.349	10.004	7.086	70,8	1.111	15,7
08	9.934	9.533	2.817	29,5	407	14,4
09	16.858	16.347	12.490	76,4	2.266	18,1
10	14.637	14.073	36	0,3	7	19,4
11	13.487	13.152	2.577	19,6	316	12,3
12	12.798	11.882	1	0,0	0	0,0
13	12.654	11.941	6.304	52,8	59	0,9
14	4.494	4.327	18	0,4	2	11,1
Totale	**194.763**	**184.878**	**113.512**	**61,4**	**18.489**	**16,3**

Il GEV di Scienze economiche e statistiche (13) ha utilizzato un algoritmo sensibilmente differente, con un diverso peso tra indicatore bibliometrico e indicatore citazionale. Innanzitutto ha utilizzato per la classificazione delle riviste non solo l'*impact factor,* ma anche l'*impact factor* a cinque anni, l'*article influcence score* e l'*h index* (calcolato per rivista e non per autore). Inoltre solo per gli articoli con un numero significativo di citazioni nelle riviste indicizzate in WoS nel periodo 2004-2010 (in rapporto agli anni trascorsi dalla pubblicazione) esse avrebbero influito sull'assegnazione della classe di merito finale. In pratica gli articoli con un numero significativo di citazioni sono stati promossi di una classe, ma gli altri non hanno subito alcuna modifica della loro classe di merito.

Un altro GEV che presenta una procedura particolare è quello delle Scienze matematiche e informatiche (1), che ha utilizzato, oltre alle banche dati WoS e Scopus anche MathSciNet[43]. Inoltre al

[43] MathSciNet, Mathematical Reviews on the web, è un database dedicato alle scienze matematiche, contiene oltre 2,8 milioni di articoli. Rende disponibili dati citazionali per riviste, autori, articoli e recensioni.

suo interno ciascun Sub-GEV ha scelto autonomamente le procedure di classificazione delle riviste anche utilizzando indicatori differenti (*impact factor*, *impact factor* a due anni, *impact factor* a cinque anni, o l'indice MQC fornito da MathSciNet).

Apparentemente il risultato migliore dal punto di vista della concordanza dei due indicatori viene ottenuto dall'Area di Scienze economiche e statistiche (13), su 6.304 prodotti valutati tramite bibliometria solo 59, meno dell'1% risultano classificati come *undecided.* Un risultato facilmente comprensibile alla luce dell'impatto ridotto dell'indicatore citazionale nella procedura adottata da questo GEV. D'altro canto il risultato meno incoraggiante è quello ottenuto per Scienze matematiche e informatiche (1); in quest'Area un articolo su quattro tra quelli valutati tramite bibliometria è stato classificato in modo discordante rispetto all'indicatore riferito all'impatto della rivista e a quello legato alle citazioni ricevute.

Capitolo 2

Obiettivi e strumenti per l'analisi metodologica della valutazione dei prodotti della ricerca nella VQR

2.1 Gli obiettivi dell'analisi

L'obiettivo generale del lavoro è l'analisi metodologica delle procedure utilizzate nel corso della VQR 2004-2010 per la valutazione dei prodotti della ricerca.

La valutazione della ricerca è immancabilmente preceduta e seguita da accesi dibattiti all'interno delle comunità scientifiche e tra queste e i decisori o gli enti che conducono la valutazione. La VQR, dalla pubblicazione del bando a quella dei suoi risultati, ha scatenato una accesa discussione dentro e fuori il mondo accademico per una serie eterogenea di ragioni. Una delle ragioni principali è identificabile nella continua ridefinizione di procedure e criteri in sede di Consiglio Direttivo e dei singoli GEV. Obiettivi, criteri e procedure valutativi vengono infatti non di rado definiti nel corso dell'esercizio «tanto da far sorgere in alcune circostanze la sensazione che la provvisorietà di alcune regole e i cambiamenti introdotti rendano la VQR simile a una sorta di *experimentum in corpore vili*» (Reale, 2013, p. 156).

Nel caso della VQR le problematiche connesse alla scelta delle procedure sono state in qualche modo amplificate dalle modalità di definizione dei criteri, di rendicontazione delle procedure stesse, di pubblicazione dei risultati. La questione dell'adeguatezza metodologica delle procedure in questo caso viene ampliata e allo stesso tempo resa più opaca a causa delle lacune nella rendicontazione dei criteri e dei metodi utilizzati.

Palumbo ha definito la valutazione come «un'attività cognitiva volta a fornire un giudizio su di un'azione (o un complesso di azioni coordinate) intenzionalmente svolta o che si intende svolgere, destinata a produrre effetti esterni, che si fonda su attività di ricerca delle scienze sociali e che segue procedure rigorose e codificabili», sottolineando ancora una volta che quest'ultimo aspetto «distingue la valutazione come impresa scientifica dalla corrente attività di espressione di un giudizio» (2001, p. 59).

La valutazione, dunque, per distinguersi da una semplice espressione di giudizio, deve attenersi a procedure «rigorose e codificabili» che rispondano ai criteri di scientificità condivisi dalla comunità scientifica. Ciò che distingue la conoscenza scientifica da altri tipi di sapere è proprio la sua riconducibilità a procedure pubbliche, riproducibili e dunque controllabili: «qualifichiamo come scientifico un insieme di asserti che sia posto, giustificato e accettato con riferimento importante, *seppure non esaustivo*, a certi criteri determinanti» (Campelli, 1999, p. 13; corsivo aggiunto).

E' evidente che nel caso della valutazione, come è in generale per le scienze sociali, gli asserti non possano fondarsi esclusivamente su dati *certi* e procedure *inattaccabili*, ed è proprio questa caratteristica epistemica a rendere fondamentale la giustificazione, o almeno l'argomentazione, delle scelte metodologiche adottate (*ivi*).

L'analisi della documentazione disponibile sulla VQR e la messa a fuoco, in particolare, dei processi di concettualizzazione, operativizzazione e rilevazione della qualità della ricerca ha evidenziato diversi punti d'ombra non di rado coincidenti con i passaggi delle procedure che maggiormente, in potenza, avrebbero potuto mettere a rischio la qualità dei dati, cioè delle valutazioni dei prodotti.

Lo studio non mira esclusivamente alla individuazione delle caratteristiche generali dell'impianto d'indagine, ma anche e soprattutto all'analisi metodologica delle soluzioni individuate in alcuni punti cruciali e dell'adeguatezza delle modalità della loro rendicontazione alla comunità.

Date le dimensioni dell'esercizio di valutazione e le sensibili differenze nelle procedure utilizzate nelle singole Aree, con riferimento ai passaggi più specifici è stato necessario limitare l'analisi a singoli casi studio. Si è scelto di fare riferimento a due sole Aree: una per la procedura di valutazione tramite peer review, l'altra per la valutazione diretta tramite analisi bibliometrica:

- l'Area 14 delle Scienze Politiche e Sociali è stata selezionata come caso-studio per la peer review, in quanto particolarmente rappresentativa dei settori non bibliometrici;
- l'Area 3 delle Scienze Chimiche è stata selezionata come caso studio per la procedura bibliometrica, poiché presenta la più alta quota di prodotti valutati tramite analisi bibliometrica (97,4%, *cfr.* Anvur, 2013a; Tab. 3.5), e che il protocollo utilizzato in quest'Area è il più frequente nella VQR (*cfr*. § 3.1).

Le questioni messe a fuoco dal punto di vista metodologico sono le tre questioni centrali nella realizzazione di ogni indagine nell'ambito delle scienze sociali: la concettualizzazione, cioè la definizione del concetto e delle sue dimensioni rilevanti, l'operativizzazione, dalla selezione degli indicatori alla costruzione di indici sintetici, e la rilevazione dei dati. Si tratta dei passaggi essenziali per la produzione di dati di qualità, "non distorti", considerando come distorto «qualsiasi dato che non soddisfi le condizioni logiche o metodologiche necessarie ai fini del conseguimento degli obiettivi cognitivi definiti a monte da chi lo ha progettato» (Mauceri, 2003, p. 41).

Il primo punto è legato alla definizione di qualità della ricerca utilizzata ai fini del giudizio finale nell'intero esercizio di valutazione, e in particolare alle sue dimensioni costitutive e alle loro relazioni reciproche. Da un punto di vista metodologico una definizione può risultare più o meno adeguata all'utilizzo in una ricerca empirica, basti pensare ai problemi legati all'ambiguità o alla vaghezza (Sartori, 1984; *cit*. in Fasanella, 1993). Sulla base dello schema proposto da Sartori è possibile analizzare il concetto di qualità della ricerca utilizzato confrontandolo innanzitutto con altre definizioni disponibili in letteratura, esaminandone poi i confini.

Nella VQR il giudizio finale di qualità per tutti i prodotti di tutte le aree si basa su tre criteri: *rilevanza*, *originalità/innovazione* e *internazionalizzazione* (Anvur, 2013a, p. 21), in questi criteri è possibile individuare le dimensioni della qualità della ricerca ritenute rilevanti ai fini della valutazione. L'analisi delle dimensioni del concetto è stata descritta da Statera come un processo insieme logico, semantico e pragmatico (Statera, 1997). Si tratta di un processo logico perché ha a che fare con i rapporti di natura *logica* tra le diverse dimensioni del concetto: rapporti che possono essere di congiunzione, disgiunzione (inclusiva o esclusiva) o negazione, implicazione o co-

implicazione. Inoltre l'analisi delle dimensioni di un concetto è un processo semantico, dato che le dimensioni di un concetto non sono che specificazioni del suo significato. Infine si tratta di un processo pragmatico «in quanto la scelta fra dimensioni che possono almeno parzialmente sovrapporsi è connessa agli scopi dell'indagine, a opzioni teoriche di fondo, e inoltre, *last but not least*, a valutazioni di maggiore o minore, più precisa o meno precisa, traducibilità operativa» (Statera, 1997, p. 129). E' sotto questi tre punti di vista (logico, semantico e pragmatico) che le dimensioni della qualità della ricerca selezionate come criteri di valutazione nella VQR sono state esaminate, soprattutto alla luce della letteratura metodologica (*cfr.* Capitolo 3).

Il secondo punto ha che fare con l'operativizzazione, strettamente intesa, del concetto di qualità della ricerca, cioè con le procedure volte a tradurre le dimensioni concettuali prima in indicatori e poi in variabili, per ricomporre infine le informazioni rilevate in un indice sintetico, e dunque tocca una serie di questioni centrali da un punto di vista metodologico. Nella VQR l'ultimo passaggio di questo insieme di operazioni è l'assegnazione di ciascun prodotto a una "classe di merito" (*Eccellente*, *Buono*, *Accettabile*, *Limitato* oppure *Non valutabile*), ma le procedure che conducono a quest'ultimo passaggio possono essere anche molto diverse tra le Aree, a volte anche tra specifici settori scientifico-disciplinari all'interno di una sola Area. Circa questa fase l'analisi metodologica deve necessariamente fare riferimento privilegiato ai due casi studio (*cfr.* Capitolo 4).

In relazione a questo punto l'interesse è diretto alla *congiunzione inespressa* (Lombardo, 1994) tra dimensioni e indicatori, un legame difficile da esplicitare se non facendo riferimento a pochi criteri, argomentabili di fronte alla comunità scientifica. Come osserva Fasanella (1993, pp. 178-9) essendo il rapporto di indicazione non necessario, dunque non esclusivo, il ricercatore lo stabilisce in base:

1. «al tipo di unità di analisi;
2. al contesto socio-culturale che include l'oggetto d'indagine […];
3. al patrimonio di conoscenze condivise dal ricercatore in quanto membro di una più ampia comunità scientifica;
4. allo scambio di opinioni, appunto circa il modo in cui si è compiuta la scelta degli indicatori, che avviene (o dovrebbe avvenire) tra il singolo ricercatore e addetti al particolare settore nell'ambito del quale egli sta indagando (v. Cartocci, 1984);
5. alla disposizione dell'indicatore ad una conversione in dati empirici ragionevolmente attendibili, valutata sulla base del confronto tra il carattere del concetto-indicatore e le tecniche di rilevazione che si intende e che è possibile impiegare».

Il quarto dei punti appena esposti, che Fasanella stesso definisce come una sorta di meta-criterio che dovrebbe basarsi sugli altri quattro (*ibidem*), viene raramente incluso nella rendicontazione e dunque nella giustificazione delle scelte operate. Questo stesso criterio riveste una particolare importanza in relazione alla valutazione della ricerca, dato che la comunità dei pari con cui dovrebbe avvenire lo scambio di opinione coincide con gli attori che è necessario coinvolgere nel processo di partecipazione/negoziazione/condivisione.

Il legame tra dimensioni e indicatori è propedeutico all'analisi metodologica della trasformazione degli indicatori in variabili, cioè in grandezze o categorie, da utilizzare per la classificazione finale dei prodotti. Il problema da analizzare è insieme di ordine semantico, logico, e procedurale e ha a che fare con la scelta del tipo di variabile in cui si trasforma l'indicatore nell'analisi dei dati. La questione sembrerebbe banale, non è affatto scontato invece l'utilizzo consapevole del

tipo di variabili scelte (Marradi, 1991b). A titolo di esempio si pensi alle variabili ordinali: la loro costruzione prevede l'assegnazione a categorie ordinate oppure l'ordinamento tramite confronto sulla base di una proprietà che si immagina come continua ma per la quale non si dispone di una adeguata unità di misura. Quest'ultima opzione comporta una serie di complicazioni, tanto che Marradi (1991b) la sconsiglia nel caso gli oggetti siano troppo numerosi. Inoltre non è semplice stimare l'ampiezza dei segmenti tra le categorie e troppo spesso nel ricorso alla serie dei numeri naturali «l'eguaglianza degli intervalli tra le categorie è solo una conseguenza meccanica, non programmata e non cercata» (Marradi, 1991b, p. 192). La situazione è ancora più complessa nel caso in cui alle categorie si attribuisca una piena autonomia semantica.

La procedura per la costruzione dell'indice finale deve tenere conto non solo del rapporto degli indicatori con il concetto e delle loro relazioni reciproche, ma anche del tipo di variabili da sintetizzare, dunque delle operazioni logiche e matematiche che è possibile o meno compiere con esse, del numero delle loro modalità o dell'ampiezza del loro campo di variazione, della loro polarità (Agnoli, 1994; Marradi, 1980 e 1991a). Nessuna operazione di sintesi può essere condotta senza una adeguata riflessione e argomentazione; come dimostrato da Nobile (2008), un uso approssimativo delle tecniche di costruzione degli indici può avere conseguenze più o meno marcate sull'accuratezza, se non addirittura sulla correttezza, dei loro risultati.

L'ultima questione è riferibile alla rilevazione di dati. Il discorso deve nuovamente connettersi specificatamente alle procedure applicate dai singoli GEV, e dunque a elementi diversi per ciascuno dei casi studio (*cfr.* Capitolo 5). L'intento qui è di considerare, dal lato della peer review, le distorsioni che possono dipendere dalla selezione dei pari e dall'assegnazione dei prodotti da valutare ai revisori, oltre che dallo strumento di rilevazione e dalle possibilità di reazione all'oggetto della valutazione; dal lato della bibliometria, le distorsioni attribuibili al tipo di database con riferimento alle sue caratteristiche. In un certo senso ci si occupa, in entrambi i casi, di quanto l'insieme delle informazioni (valutazioni) rilevate possa dirsi non distorto sulla base di *chi* (o *cosa*) le ha fornite, mentre analizzando le definizioni operative si mette a fuoco *come* le informazioni sono state richieste/rilevate e processate.

In relazione alla peer review le prime questioni da analizzare sono legate alle modalità di selezione dei revisori e al loro esito, in relazione sia ai requisiti di trasparenza e pubblicità della procedura che alle possibili distorsioni legate alle caratteristiche (al campo si studi, all'età, al genere, al prestigio e al ruolo accademico) di revisori e autori (per una rassegna recente si veda Lee *et al.* 2013). In secondo luogo vanno considerate le distorsioni che potrebbero intervenire a causa del *matching* tra revisori e prodotti, non solo in ragione delle caratteristiche di entrambi, ma anche in relazione ad asimmetrie nel carico di lavoro, differenze nella varietà dei prodotti per i singoli revisori, o nella varietà dei revisori per i prodotti di uno stesso autore. In relazione a questi e altri fattori vanno considerate le possibili conseguenze di una valutazione che non preveda l'oscuramento né degli autori dei prodotti né della loro collocazione editoriale.

Infine un discorso a sé andrebbe condotto sull'omogeneità e la stabilità delle scale di giudizio utilizzate dai revisori (Callaham *et al.* 1998), nonostante l'impossibilità di analizzare approfonditamente questo aspetto sulla base dei dati resi pubblici dall'Anvur.

Con riferimento all'analisi bibliometrica, le problematiche metodologiche da affrontare sono sostanzialmente quelle proprie di tutte le analisi secondarie. Non essendo la base di dati costruita *ad hoc*, questi potrebbero non essere di qualità oppure non rispondere completamente alle necessità

informative del ricercatore, limitando dunque gli obiettivi cognitivi dell'indagine oppure distorcendone i risultati. Verranno dunque analizzati i metadati disponibili sui database utilizzati (per l'Area 3 WoS e Scopus), e gli algoritmi di calcolo degli indici, conteggio delle citazioni, eccetera. La scelta del database non è indifferente in ambito scientometrico. Non solo le basi di dati non sono identiche, ma anche gli algoritmi e gli indici calcolati sono diversi; se i risultati sono altamente correlati tra i database a livello di Paese, per singoli campi disciplinari non sempre i risultati ottenuti da database diversi sono simili (Adam, 2002; Bakkalbasi *et al.* 2006; Bar-llan, 2008; Archambault *et al.* 2008; Archambault *et al.* 2009; Biocati-Rinaldi, 2010). Alla luce dei metadati e degli studi disponibili si esamineranno le scelte del GEV di Scienze Chimiche in riferimento ai database e agli indici utilizzati.

2.2 Gli strumenti dell'analisi

Il materiale empirico disponibile sui criteri e le procedure utilizzate per la valutazione dei prodotti nel corso della VQR è composto soprattutto dai documenti ufficiali che istituiscono, descrivono o rendicontano l'esercizio di valutazione, e dai metadati relativi ai database citazionali utilizzati. Questi documenti includono il bando (Anvur, 2011) e le tre parti del rapporto finale: *Statistiche e risultati di compendio* (Anvur, 2013a), *La valutazione delle singole strutture* (Anvur, 2013b) e *I confronti internazionali per le aree bibliometriche* (Anvur, 2013c), incluse naturalmente le relative appendici. Inoltre in relazione alle due aree selezionate come casi-studio verranno analizzati approfonditamente i report di Area, le relative appendici e le tavole pubblicate dall'Anvur (Anvur, 2013d[44]; Anvur, 2013d, GEV 3; Anvur, 2013d, GEV14).

Il disegno d'indagine alla base del progetto di tesi prevede l'utilizzo dei documenti ufficiali come fonte primaria di informazione, di conseguenza è centrato sostanzialmente sull'analisi documentaria. L'analisi dei testi e delle tavole pubblicati dall'Anvur e dai singoli GEV ha permesso di individuare e classificare i segmenti della documentazione utili per la ricostruzione e l'analisi della valutazione dei prodotti nel corso dell'esercizio.

Una volta ricostruite adeguatamente le procedure e individuati i passaggi decisivi per la qualità dei dati prodotti, la strategia di ricerca ha previsto essenzialmente il confronto delle procedure con gli standard metodologici attualmente condivisi nel campo della ricerca sociale e della bibliometria, con un ampio uso della letteratura internazionale sul tema della valutazione della ricerca in generale e sugli strumenti utilizzati nel corso della VQR in particolare.

L'analisi dei documenti è stata seguita da una serie di interviste focalizzate, dirette a testimoni privilegiati (alcuni componenti del Consiglio Direttivo dell'Anvur, i presidenti e i coordinatori dei GEV delle Aree 3 e 14), con lo scopo primario di approfondire alcuni passaggi delle procedure rimasti in ombra nella documentazione ufficiale, ma a questo obiettivo si aggiunge l'intenzione di mettere a fuoco le opinioni degli EV sui criteri e procedure, sulla loro affidabilità e sul loro (eventuale) impatto sulle comunità scientifiche.

[44] I riferimenti ai report di Area (Anvur, 2013d) nel testo saranno sempre seguiti dal GEV di riferimento dell'Area; ad esempio per l'Area 3: Anvur, 2013d, GEV 3; per l'Area 14: Anvur, 2013d, GEV 14.

2.2.1 Le interviste focalizzate

L'analisi dei documenti, ha messo in evidenza la mancanza di trasparenza o chiarezza della documentazione ufficiale e nei rapporti in riferimento ad alcuni passaggi della procedura di valutazione dei prodotti. Mediante le interviste si è, dunque, tentato innanzitutto di integrare le informazioni disponibili nei documenti ufficiali circa le procedure di valutazione dei prodotti. Diversi passaggi (ad esempio, nella procedura in peer review, l'attribuzione dei prodotti ai revisori; nella procedura bibliometrica, la scelta della *subject category* in cui valutare il prodotto; in entrambe le modalità decisionali nell'ambito dei gruppi di consenso; *cfr.* Capitolo 3) sono infatti esposti poco chiaramente, o con eccessiva sintesi, nei rapporti. Accanto a questo primo obiettivo, strettamente legato al piano informativo, vi sono diversi obiettivi cognitivi legati all'analisi metodologica svolta, dunque alle opinioni degli esperti valutatori e dei componenti Anvur circa la qualità del dato nell'esercizio di valutazione della ricerca.

Innanzitutto le interviste mirano a ricostruire i tratti essenziali della valutazione della ricerca dal punto di vista dei testimoni, e di rilevare l'importanza che la qualità del dato assume in relazione a questi quadri specifici. Si cerca di chiarire quanto e come nel corso dell'esercizio di valutazione si sia posta attenzione alla questione della qualità dei dati e di rilevare le opinioni dei testimoni, per poi, dopo aver discusso i vari passaggi della procedura, mettere a fuoco dal punto di vista dell'intervistato i passaggi che sarebbe preferibile migliorare, non solo in riferimento alla qualità del dato, ma anche in un'ottica più ampia.

Un ulteriore obiettivo è il confronto con i testimoni circa i criteri di valutazione e le classi di merito utilizzati nel corso della VQR. Risulta infatti di grande interesse l'opinione di chi ha progettato o attuato le procedure circa i presupposti della valutazione, fissati a monte da un decreto ministeriale (DM 17 del 15 luglio 2011, art. 8, comma 2) e ripresi dal bando di partecipazione all'esercizio (Anvur, 2011, p. 7). Nella stessa direzione si colloca l'approfondimento del ruolo dell'Anvur e dei GEV nella definizione delle procedure e degli strumenti utilizzati per valutare i prodotti, oltre che nella gestione della procedura e dell'eventuale coinvolgimento di altri attori (società scientifiche, atenei, dipartimenti, editori) nella definizione o nell'implementazione delle procedure, nell'ottica di una validazione della valutazione tramite la partecipazione, negoziazione e condivisione con gli attori coinvolti (Fasanella, 2013).

Un obbiettivo cognitivo aggiuntivo, meno legato all'analisi metodologica delle procedure e più vicino all'ambito della sociologia della scienza, è riferito alle opinioni e alle aspettative degli esperti valutatori circa l'impatto della VQR sulle comunità scientifiche.

Nella letteratura internazionale non mancano gli esempi di mutamenti nelle modalità di comunicazione dei risultati della ricerca a seguito dell'implementazione di specifiche procedure di valutazione (Westerheijden, 1997; Georghiou *et al.* 2000; Talib, 2001; Gläser *et al.* 2002; Butler, 2003a; Geuna e Martin, 2003; Jimenez-Contreas *et al.* 2003 ; Moed, 2008; Osuna *et al.* 2010), ed esiste una letteratura dedicata apertamente alle "conseguenze indesiderate" dell'utilizzo di indicatori bibliometrici (per tutti Weingart, 2005; Burrows, 2012). L'obiettivo di questa sezione dell'intervista non è la semplice rilevazione delle aspettative dei valutatori circa le future tendenze nella produzione scientifica, ma anche la discussione del legame tra i criteri, le procedure e l'impatto del sistema di valutazione. L'intento è di mettere in luce non solo l'intenzionalità o il carattere inatteso delle eventuali conseguenze, ma anche l'opinione dei singoli intervistati circa la loro desiderabilità.

2.2.1.1 La tecnica di intervista

Le interviste sono dirette ai testimoni privilegiati dell'esercizio di valutazione, dunque a soggetti esperti in grado di integrare ed ampliare nel corso dell'intervista non solo le informazioni disponibili per l'analisi, ma anche i suoi obiettivi, fornendo interpretazioni e chiavi di lettura. Le tracce delle interviste sono scarsamente strutturate e lo stile di conduzione estremamente flessibile proprio per permettere ai testimoni di andare oltre le informazioni richieste, creando un clima di confronto costruttivo sulla cui base discutere le procedure e i loro esiti.

Dal punto di vista del grado di standardizzazione si tratta di interviste semi-standardizzate: «il protocollo di rilevazione, oltre all'elenco dei temi da trattare, prevede anche delle domande da porre all'intervistato. Queste domande non contemplano alternative di risposta prefigurate e possono essere adattate di volta in volta, sia nella formulazione sia nell'ordine, alle caratteristiche del soggetto che si sta intervistando» (Mauceri, 2003, p. 34). La ragione per cui si è scelto questo grado di standardizzazione è legata sia alle caratteristiche degli intervistati che agli obiettivi cognitivi dell'intervista: «questo modo di condurre l'intervista concede ampia libertà ad intervistato e intervistatore, garantendo nello stesso tempo che tutti i temi rilevanti siano discussi e che tutte le informazioni necessarie siano raccolte» (Corbetta, 1999, p. 415).

In particolare le interviste sono state progettate come interviste focalizzate (Merton, Fiske e Kendall, 1956). Questa tecnica di intervista, messa a punto nell'ambito della ricerca sugli effetti e le dinamiche della comunicazione di massa, si differenzia dall'intervista in profondità essenzialmente perché è destinata a soggetti coinvolti in una stessa esperienza/situazione e prevede un'approfondita conoscenza da parte del ricercatore della esperienza/situazione oggetto di studio, mirando alla messa a fuoco di informazioni, opinioni e aspettative ad essa legate. Naturalmente il fatto che gli intervistati siano in questo caso a tutti gli effetti testimoni privilegiati, in quanto esperti di valutazione coinvolti attivamente in prima persona in un esercizio di valutazione, rende particolarmente adatto l'uso di uno strumento mirato a mettere a fuoco specifiche informazioni, oltre che le opinioni e aspettative rilevanti in relazione agli obiettivi cognitivi dello studio.

L'utilizzo di strumenti semi-standardizzati presenta naturalmente degli svantaggi con riferimento alla completezza e alla comparabilità delle informazioni rilevate, tuttavia il semplice fatto che le interviste siano condotte in prima persona da chi le ha progettate dovrebbe ridurre i rischi relativi all'insufficienza o all'irrilevanza delle informazioni raccolte ai fini del raggiungimento degli obiettivi cognitivi. Inoltre non è la comparabilità delle informazioni quanto la loro varietà e completezza a risultare fondamentale per la riuscita dello studio.

Infine, non è stata utilizzata un'unica traccia di intervista per tutti gli intervistati. Se infatti alcune aree tematiche da mettere a fuoco e alcuni quesiti cui dare risposta riguardano tutti i testimoni, altri elementi sono diversi a seconda del ruolo svolto dall'intervistato nel corso dell'esercizio di valutazione.

2.2.1.2 I testimoni privilegiati

Le interviste si configurano come interviste a testimoni privilegiati in quanto dirette a «persone particolarmente qualificate a rispondere su determinate questioni» (Madge, 1962, tr. it. 1996, p.

189) che hanno «seguito una data situazione "dall'interno", e quindi possono fornire informazioni "di prima mano" di notevole interesse» (Guala, 2000, p. 359).

I testimoni selezionati per queste interviste sono componenti del Consiglio Direttivo dell'Anvur o membri dei GEV delle Aree selezionate come casi-studio, privilegiati non solo grazie all'esperienza specifica nell'ambito della VQR, ma anche per l'esperienza più generale nel campo della valutazione della ricerca. In ragione del numero elevato di membri dei due GEV: 23 per l'Area 3 e 13 per l'Area 14 (Tabella 9), si è resa necessaria una selezione dei testimoni da intervistare: la natura poco strutturata e focalizzata delle interviste ha suggerito di limitarne il numero perché l'analisi potesse risultare efficace e sufficientemente approfondita. Una sola delle interviste previste non è stata in effetti realizzata.

Tabella 9 – Composizione numerica del Consiglio Direttivo e dei due GEV delle Aree selezionate come casi-studio e numero di interviste progettate

	Membri	**Interviste previste**	**Interviste realizzate**
Consiglio Direttivo Anvur	7	3+1	3+1
Gev 3	23	4+1	3+1
Gev 14	13	3+1	3+1
Totale	***43***	***13***	***12***

La selezione (il cui esito è presentato in Tabella 10) è stata basata essenzialmente sul ruolo dei soggetti, favorendo sia all'interno dei GEV che nel Consiglio Direttivo gli Esperti che hanno svolto un ruolo di coordinamento e che dunque hanno avuto nel corso dell'esercizio una posizione privilegiata di osservazione in riferimento alle procedure. Nel caso dell'Anvur sono stati selezionati il Presidente, il coordinatore ed il vice-coordinatore della VQR ed il Direttore, mentre per entrambi i GEV delle Aree selezionate come casi-studio le interviste si sono limitate ai Presidenti e ai coordinatori dei sub-GEV. Inoltre nel corso delle interviste è emersa l'opportunità di sentire anche testimoni con altri ruoli, dunque sono stati contattati e intervistati, in aggiunta ai primi, anche soggetti che nel corso della VQR hanno avuto funzioni gestionali e organizzative: il Direttore dell'Anvur e le due Assistenti-GEV delle Aree selezionate come casi studio.

I contatti per le interviste sono iniziati nel mese di Ottobre del 2014, tuttavia a causa di problemi organizzativi e comunicativi le interviste si sono svolte in un arco temporale più ampio di quanto auspicato: tra il mese di Novembre 2014 e il Marzo 2015. Tutte le interviste inizialmente pianificate o aggiunte in itinere sono state realizzate, ad eccezione dell'intervista con il professor Maurizio Prato (Coordinatore sub-Gev di Chimica Organica e Medica, in grigio in Tabella 10) con il quale non è stato possibile fissare un incontro. Inoltre per questioni organizzative non è stato possibile intervistare individualmente il dottor Torrini (Direttore dell'Agenzia) e la dottoressa Blasi (assistente GEV di Area 14), le interviste con questi due rispondenti si sono dunque svolte simultaneamente.

Tabella 10 – Componenti del Consiglio Direttivo e membri dei GEV selezionati come testimoni privilegiati

Ambito	Ruolo	Nominativo	Afferenza
Anvur	Presidente	Fantoni Stefano	
	Componente (Coordinatore VQR)	Sergio Benedetto	
	Componente (Vice-coordinatore VQR)	Bonaccorsi Andrea	
	Direttore	Roberto Torrini	
GEV 3	Presidente	Barone Vincenzo	Scuola Normale Superiore di Pisa
	Coordinatore sub-GEV Chimica inorganica e industriale	Pacchioni Gianfranco	Universita'degli Studi di Milano-Bicocca
	Coordinatore sub-GEV Chimica organica e medica	Prato Maurizio	Universita'degli Studi di Trieste
	Coordinatore sub-GEV Chimica analitica e fisica	Torsi Luisa	Universita'degli Studi di Bari Aldo Moro
	Assistente-GEV	Valentina Carletti	
GEV 14	Presidente	Colozzi Ivo	Universita'degli Studi di Bologna
	Coordinatore sub-GEV Scienze Politiche	Bazzicalupo Laura	Universita'degli Studi di Salerno
	Coordinatore sub-GEV Scienze Sociali	Cipriani Roberto	Universita'degli Studi Roma Tre
	Assistente-GEV	Brigida Blasi	

E' il caso di sottolineare che alcune informazioni circa la procedura non sono risultate rilevabili neppure tramite le interviste, dato che i componenti dei GEV non erano informati oppure non avevano memoria di alcuni passaggi tecnici, a volte di non secondaria importanza (ad esempio circa il calcolo delle distribuzioni cumulate per la procedura bibliometrica). Allo scopo di chiarire questi passaggi specifici sono stati contattati il dottor Alberto Francesco Anfossi, Assistente-GEV di Area 2, che ha seguito in prima persona la messa a punto del sistema di assegnazione delle classi di merito utilizzata in quasi tutte le Aree per la procedura di valutazione bibliometrica e la dottoressa Giovanna Colizza, Assistente GEV di Area 14 durante le prime fasi della VQR.

2.2.1.3 Le tracce di intervista

Sono state progettate e utilizzate tre diverse tracce di intervista, calibrate in base al ruolo svolto dagli intervistati nel corso della VQR e mirate a rilevare informazioni differenti o a un differente grado di specificità. La prima traccia, più generale, è indirizzata ai membri del Consiglio Direttivo dell'Anvur (Traccia di intervista Anvur, Appendice A), la seconda, centrata sulla procedura di valutazione diretta tramite analisi bibliometrica, è destinata ai membri del GEV 3, Area delle Scienze Chimiche (Traccia di intervista Area 3, Appendice A), la terza, centrata sulla procedura di valutazione tramite peer review è concepita per i membri del GEV 14, Area delle Scienze Politiche e Sociali (Traccia di intervista Area 14, Appendice A).

Alcune aree tematiche sono comuni a tutte e tre le tracce, altre sono specificamente pensate per i membri del Consiglio Direttivo o per gli Esperti Valutatori o direttamente connesse a una delle due procedure utilizzate per la valutazione dei prodotti, ma nell'insieme la struttura delle tracce risulta abbastanza uniforme (Tabella 11). Le aree tematiche sostanzialmente identiche nelle tre tracce sono riferibili alla concezione della valutazione e della qualità del dato, e alle opinioni sugli eventuali impatti della VQR sulle comunità scientifiche.

Tabella 11 – Aree tematiche per ciascuna traccia di intervista

Consiglio Direttivo Anvur	*GEV 3*	*GEV 14*
Valutazione e qualità dei dati	Valutazione e qualità dei dati	Valutazione e qualità dei dati
Le procedure	La procedura bibliometrica	La procedura di peer review
	I gruppi di consenso	I gruppi di consenso
		I referaggi aggiuntivi
Opinioni e suggerimenti sulla procedura di valutazione	Opinioni e suggerimenti sulla procedura di valutazione	Opinioni e suggerimenti sulla procedura di valutazione
Opinioni e suggerimenti sui criteri di valutazione	Opinioni e suggerimenti sui criteri di valutazione	Opinioni e suggerimenti sui criteri di valutazione
Il ruolo del Consiglio Direttivo nella procedura di valutazione		
Il ruolo dei GEV nella procedura di valutazione	Il ruolo del GEV nella procedura di valutazione	Il ruolo del GEV nella procedura di valutazione
Opinioni sull'impatto della VQR	Opinioni sull'impatto della VQR	Opinioni sull'impatto della VQR

Le tracce destinate agli Esperti Valutatori risultano più dettagliate in riferimento alle procedure di valutazione dei prodotti, mentre la traccia per i componenti del Consiglio Direttivo è maggiormente centrata sul ruolo svolto dal Consiglio e dai GEV nel corso dell'esercizio di valutazione. Entrambe contengono diverse domande circa il coinvolgimento di altri attori (società scientifiche, atenei, dipartimenti, editori), nelle varie fasi di definizione ed implementazione dell'esercizio di valutazione. I quesiti riferiti alle procedure sono molto diversi nelle due tracce destinate ai GEV, proprio in ragione delle differenze tra le due procedure utilizzate (*cfr.* Appendice A). Alcune aree tematiche risultano estremamente simili, seppure calibrate sul ruolo dell'intervistato, in particolare le due aree riferibili alle opinioni sulla procedura di valutazione e sui criteri, oltre all'area riferita al ruolo dei GEV nella procedura di valutazione.

2.2.1.4 L'analisi e la restituzione dei contenuti delle interviste

La trascrizione delle interviste è accurata, ma non testuale, per permettere una maggiore leggibilità[45], lasciando inalterato il lessico ed i significati espressi dagli intervistati.

Data la doppia natura, informativa ed esplorativa, degli obiettivi delle interviste, si è optato per una tecnica di analisi in parte ispirata all'approccio proposto dai sostenitori della *Grounded Theory* (Glaser e Strauss, 1967). Evidentemente l'analisi qui non mira alla costruzione di una teoria, tuttavia proprio in ragione del duplice obiettivo delle interviste, è indiscutibile il vantaggio offerto dall'adozione di un'ottica in cui la raccolta dei dati, l'individuazione delle categorie in cui organizzarli e la loro analisi sono strettamente interconnesse e si sviluppano in parallelo.

Nei termini di Demazière e Dubar (2000) l'atteggiamento di analisi adottato potrebbe essere denotato principalmente come un atteggiamento illustrativo: «le parole degli intervistati sono utilizzate al fine di "illustrare" le affermazioni del ricercatore» (Bichi, 2002, p. 148), ed in secondo

[45] Ad esempio nella trascrizione non compaiono le ripetizioni tipiche del linguaggio parlato, così come gli intercalari; esitazioni e bravi pause sono state riportate nel testo (simbolizzate da tre puntini), mentre le note sono state limitate al minimo indispensabile: interruzioni, elementi di disturbo, azioni rilevanti, identificazione di altri soggetti citati nel corso dell'intervista, qualora rilevanti.

luogo come un atteggiamento restitutivo, in cui: «la parola dell'intervistato viene considerata trasparente, in grado di fornire da sola i significati utili alla comprensione» (*ibidem*, p. 149). Infatti, in relazione all'obiettivo informativo delle interviste, l'analisi è stata sviluppata attraverso la classificazione di stralci di intervista in categorie e sotto-categorie e la restituzione dei contenuti delle interviste avviene, sostanzialmente, riportando nelle note a piè di pagina i passaggi utili alla precisazione o alla interpretazione delle procedure utilizzate per la valutazione dei prodotti (Capitoli 4, 5 e 6). In relazione all'obiettivo esplorativo, connesso essenzialmente agli impatti della VQR, l'analisi è andata oltre la classificazione degli stralci, mirando all'individuazione delle relazioni tra gli obiettivi della VQR, gli impatti attesi ed i possibili impatti indesiderati (Capitolo 7). La presentazione dei risultati non riporta integralmente il testo delle interviste, come un puro atteggiamento restituivo prevedrebbe, ma avviene attraverso l'individuazione degli stralci rilevanti, in cui le parole degli intervistati hanno un valore interpretativo oltre che informativo. Questi stralci sono riportati direttamente nel testo e costituiscono l'ossatura del discorso.

Il principale strumento di organizzazione ed analisi della base empirica utilizzato è il software Atlas.it[46], che permette di gestire contemporaneamente più documenti, effettuando le analisi sui singoli documenti (in questo caso le singole interviste, *primary documents*[47]), su sezioni specifiche dei documenti (stralci, *quotations*), e sull'intero *corpus* (cioè sull'insieme delle interviste, *hermeneutic unit*). Il programma opera su una unità, il codice (*code*), che viene definita dal ricercatore e può essere memorizzata in riferimento a più segmenti di testo (*quotations*)[48]. L'analisi mira essenzialmente alla costruzione delle categorie tematiche, cioè dei codici, e all'individuazione delle relazioni rilevanti sulla base dell'evidenza empirica costituita dal testo.

L'utilità di questo approccio per l'individuazione di specifiche informazioni nelle interviste è evidente, il programma permette infatti di interrogare il testo richiamando stralci di testo sulla base del codice (o dei codici) ad essi associati, sull'intero *corpus* o su parte di esso[49].

In relazione ai codici connessi con gli impatti dell'esercizio il programma ha permesso di costruire dei *network* categoriali, cioè reti in cui i codici costituiscono i nodi e le relazioni tra queste sono presentati da frecce differenziate in base al tipo specifico di relazione (appartenenza, associazione, ecc.). E' importante sottolineare che, così come la categorizzazione e l'individuazione dei codici, gli elementi da includere nella rete, le relazioni e le loro connotazioni sono frutto esclusivamente del lavoro di interpretazione: il programma permette di offrirne una rappresentazione semplice ed intuitiva, ma non contribuisce alla costruzione dell'impianto.

[46] Per una trattazione più esaustiva delle caratteristiche e delle opzioni offerte dal programma di analisi si rimanda a Bichi, 2002.

[47] L'intervista simultanea al direttore Anvur Torrini e alla dottoressa Blasi è stata importata due volte, in modo tale da permettere la classificazione distinta degli stralci riferibili a ciascun intervistato.

[48] Documenti, codici e stralci possono essere raggruppati in categorie, denominate famiglie (*families*).

[49] Nel corpus ciascuna intervista costituiva un documento primario, ma questi erano organizzati in tre famiglie:Anvur, Area 14, Area 3, in base al gruppo di appartenenza degli intervistati.

Capitolo 3
La qualità della ricerca

Introduzione

Nella ricerca sociale la definizione del concetto da rilevare è un momento essenziale non solo dal punto di vista teorico, ma anche dal punto di vista metodologico. Nel caso in cui l'indagine abbia una finalità valutativa, bisognerebbe utilizzare la massima accortezza e trasparenza possibili in questa fase, non solo per le esigenze di *accountability* già poste in evidenza, ma anche e soprattutto per permettere la realizzazione di una negoziazione degli obiettivi dell'indagine con i diversi *stakeholders*.

Il concetto centrale nella VQR è senza alcun dubbio quello di qualità della ricerca, e nell'adeguatezza della sua definizione va ricercato il senso stesso dell'esercizio di valutazione. Per poter procedere all'analisi metodologica della procedura di valutazione dei prodotti nella VQR è dunque necessario innanzitutto presentare e analizzare criticamente la definizione di qualità della ricerca utilizzata ai fini del giudizio finale nell'intero esercizio di valutazione, con particolare riferimento alle sue dimensioni costitutive e alle loro relazioni reciproche.

3.1 Il concetto di qualità della ricerca

I concetti, nella formulazione di Marradi, sono ritagli operati in un flusso di esperienze infinito in estensione e profondità e infinitamente mutevole (1980, p. 9). Questa concezione risulta non solo decisamente weberiana, ma anche molto vicina alla visione di Lazarsfeld, che assegnava alla rappresentazione figurata del concetto una funzione generativa dei concetti sociologici, individuando nell'osservazione dei fenomeni sociali il punto di partenza del processo di formazione dei concetti stessi, dato che un concetto «corrisponde ad un insieme complesso di fenomeni, piuttosto che ad un fenomeno semplice e facilmente osservabile» (Lazarsfeld 1958, p. 43).

La maniera in cui il "ritaglio" cui Marradi fa riferimento viene operato: «non è dettata in forma cogente da qualità intrinseche delle nostre sensazioni (o dalle cose in sé, come pensavano i filosofi scolastici), ma dipende in larga misura dalle necessità pratiche di un certo individuo, gruppo, società, ecc.» (1980, p. 10). E' dunque fondamentale partire dall'assunto che i concetti non sono pensabili come veri o falsi (Marradi, 1980; Fasanella, 1993), ma solo come utili o inutili ai fini per cui sono preposti, che questi siano legati alla comunicazione nella vita quotidiana, o alla formulazione di teorie e ipotesi nell'ambito scientifico (Marradi, 1980). L'utilità di un concetto è inoltre pensabile nell'ambito della conoscenza scientifica come funzione da un lato del loro grado di astrazione e dall'altro dalla relazione che può essere stabilita tra essi e dei fatti osservabili (Fasanella, 1993).

Nel corso della trattazione dunque non viene proposta alcuna definizione di qualità della ricerca alternativa a quella proposta dall'Anvur, ma ne sono presi in considerazione gli aspetti

fondamentali, nel tentativo di analizzarli dal punto di vista dell'utilità in relazione agli obiettivi sostantivi dell'esercizio di valutazione, ma anche con riferimento alla traducibilità in termini empirici.

La questione della definizione dei concetti nella ricerca scientifica è molto complessa, e il dibattito in proposito ha riguardato (e riguarda) diversi nodi cruciali: primo fra tutti se sia il concetto o il(i) termine(i) che lo designa ad essere definito. Marradi (1980) include tra chi propende per la definizione del concetto studiosi come Boudon e Lazarsfeld (1965) e Nowak (1976), tra coloro che invece propendono per la definizione dei termini annovera tra gli altri Hempel (1952) e Carnap (1938), ma evidenzia che è il rapporto tra concetto e termine(i) a essere definito. Infatti «se il *definiens* (la frase che viene dopo il segno di equivalenza) è un termine solo (come in stato = situazione), si dichiara che tale raccordo è simile a un raccordo tra lo stesso concetto e un altro termine. Se, come nella maggior parte dei casi, il *definiens* è costituito da più termini che rimandano a più concetti, si dichiara prima un'equivalenza semantica fra il concetto da definire e una determinata combinazione logica di altri concetti, indi una conseguente equivalenza tra il termine che designa il primo concetto e una data costruzione sintattica dei termini che designano gli altri» (Marradi, 1980, p. 18).

Esistono inoltre diversi tipi di definizioni, ciascuno dei quali assume una funzione specifica. La definizione della qualità della ricerca nella VQR si connota come una definizione *stipulativa*[50], nel senso che propone un nuovo raccordo tra il concetto e i termini che lo designano. Si tratta di una definizione in cui «sia il concetto sia il termine possono essere proposti ex novo oppure essere già in uso: quello che conta è che il raccordo tra loro sia nuovo, e sia proposto alla comunità come utile a fini pratici o cognitivi» (*ibidem*). Non vi è, infatti, nulla di nuovo nella locuzione "qualità della ricerca", ciò che risulta importante è definire il concetto in relazione al fine pratico e cognitivo della valutazione, specificando il significato che assume nel contesto in cui viene utilizzato: la VQR.

Il concetto di qualità presenta, infatti, intrinsecamente, un certo grado di ambiguità e relatività[51] (Frudà, 1997) che rende necessaria l'individuazione di criteri di riferimento. Se invero la qualità in sé è un concetto a un altissimo livello di astrazione e generalità, il riferimento alla ricerca ne diminuisce il grado di generalità, ma non quello di astrazione (Marradi, 1980). Perché sia possibile ridurre il grado di astrazione del concetto è necessario fare riferimento a concetti più specifici, cioè la definizione deve risultare in grado di stipulare il significato attribuito al concetto legandolo a un insieme di significati meno astratti: in questo caso i criteri di valutazione.

Una definizione, per quanto stipulativa e non ancora operativa, può risultare più o meno adeguata all'utilizzo in una ricerca empirica. Tra le questioni da considerare e controllare vi sono innanzi tutto i rischi legati all'ambiguità o alla vaghezza della definizione (Sartori, 1984; *cit*. in Fasanella, 1993). Nell'ottica di Sartori l'ambiguità è intesa come una proprietà della relazione tra il concetto e il(i) termine(i) che lo indicano, mentre la vaghezza si connota come una proprietà della relazione tra il concetto e i suoi i referenti empirici (*ibidem*).

Sartori indica due passaggi necessari a evitare tanto l'ambiguità quanto la vaghezza della definizione: il primo consiste nello stabilire una definizione *dichiarativa*, stabilendo univocamente la

[50] Questo genere di definizione viene denominato "definizione nominale" da Hempel (1952) e Pasquinelli (1977); "definizione esplicativa" da Carnap (1936).

[51] Si pensi, ad esempio, all'ampio dibattito scaturito dal tentativo di definire, stipuativamente e poi operativamente, il concetto di "qualità della vita"; per una rassegna si vedano: Rapley, 2001, Ferriss, 2004; Sirgy et al., 2006.

relazione tra termine(i) e significato (concetto); il secondo consta nella definizione *denotativa*, a sua volta scomponibile in *precisante*, *operativa* e *ostensiva*, mirata a circoscrivere i referenti empirici del concetto (Sartori, 1984; Fasanella, 1993; Lombardo, 1994). La definizione precisante, di particolare interesse in questa sede[52], consente di passare dal concetto generale a uno o più concetti specifici, ciascuno dei quali deve a sua volta essere definito attraverso una definizione dichiarativa (Lombardo, 1994). Nel caso del concetto di qualità della ricerca nella VQR non siamo di fronte a definizione in senso stretto (cioè a una definizione in cui è data una equivalenza logica tra *definiendum* e *definiens*), ma a una specificazione del significato, che consiste: «nella descrizione via indicatori (*indicators*) e riferimenti (*references*) delle condizioni sotto le quali è opportuno applicare il termine. Tale specificazione è comunque incompleta, in quanto ulteriori indicatori potrebbero essere aggiunti ed alcuni altri potrebbero essere sostituiti; nondimeno gli indicatori stipulati hanno uno status logico – ancorché contestuale e funzionale – simile a un *definiens*: la loro relazione col termine di cui è specificato il significato è materia di decisione (Kaplan, 1955, p. 530)» (Agnoli, 1999, p. 207).

I criteri alla base della valutazione nella VQR, dunque le dimensioni della qualità della ricerca, previste dal Decreto Ministeriale che ha dato l'avvio alla procedura di valutazione (DM 17 del 15 luglio 2011) e dal bando dell'Anvur (2011, p. 7), sono:

a) la *rilevanza*: «da intendersi come 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» (Anvur, 2011, p. 7);
b) l'*originalità/innovazione*: «da intendersi come contributo all'avanzamento di conoscenze o a nuove acquisizioni nel settore di riferimento» (*ibidem*);
c) l'*internazionalizzazione*: «da intendersi come posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica, inclusa la collaborazione esplicita con ricercatori e gruppi di ricerca di altre nazioni» (*ibidem*).

Relativamente ai brevetti, un ulteriore criterio (d) era riferito «*al trasferimento*, *allo sviluppo tecnologico e alle ricadute socio-economiche (anche potenziali)*» (*ibidem*).

Facendo riferimento alla classificazione finale dei prodotti nella VQR è possibile risalire a quanto ciascuno di questi criteri contribuisce alla definizione della qualità della ricerca. L'obiettivo finale della valutazione è infatti l'assegnazione di ciascun prodotto a una *classe di merito*, in una classificazione che fa riferimento alla collocazione delle pubblicazioni nella «scala di valore condivisa a livello internazionale» (Anvur, 2011a, p. 7). Il bando del 2011 riporta le seguenti definizioni lessicali delle quattro classi di merito:

«1. 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.
5. I prodotti di livello *buono* sono quelli di importanza internazionale e nazionale riconosciute per originalità dei risultati e rigore metodologico.

[52] La definizione operativa è trattata diffusamente nel prossimo capitolo, mentre circa la definizione ostensiva, cioè dell'atto linguistico di definire un oggetto mostrandolo all'interlocutore, ci si limita a osservare con Lombardo che il suo utilizzo risulta «estremamente limitato, se non inesistente, nelle scienze sociali» (Lombardo, 1994, p. 23).

6. 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.
7. 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» (*ibidem*).

Le questioni relative alle relazioni tra le dimensioni della qualità della ricerca saranno approfondite nel prossimo paragrafo, tuttavia è importante sottolineare che il concetto in analisi viene presentato come un concetto strutturato (Di Giammaria, 2009; Fasanella, 2010), le cui dimensioni costitutive presentano determinate relazioni e non altre.

La formulazione dei criteri sulla base dei quali valutare la qualità della ricerca si configura dunque come una specificazione del significato, frutto di una decisione circa la loro relazione col concetto, che descrive, attraverso una serie di riferimenti e indicatori (nel senso massimamente ampio del termine), le condizioni per cui a un dato oggetto (prodotto) può essere attribuita una maggiore o minore rispondenza a ciascuno dei criteri. Evidentemente si tratta di una definizione incompleta, riconducibile all'alveo della riduzione più che a quello della definizione vera e propria (Agnoli, 1994)[53]. La definizione della qualità della ricerca si configura come una riduzione perché evidenzia solo gli aspetti del concetto ritenuti rilevanti ai fini dell'esercizio di valutazione, ma soprattutto perché è legata a una proprietà che in astratto è in grado di assumere un numero di valori potenzialmente infinito, e pertanto non precisamente rilevabile.

Tralasciando la caratteristica di incompletezza della definizione, del resto riferibile alla massima parte dei concetti utilizzati nelle scienze sociali, quella di qualità della ricerca nei termini di Carnap si configurerebbe come una definizione *esplicativa*, che mira ad assegnare a espressioni dal significato più o meno vago significati precisamente determinati. Agnoli sottolinea che «muovendo da significati usuali l'esplicazione mira dunque ad attenuare ambiguità, limitazioni e contraddizioni nell'uso» (1994, p. 89).

Tornando dunque alle questioni dell'*ambiguità* e della *vaghezza*, è evidente che in questo caso queste due proprietà vadano riferite ai singoli criteri prima che al concetto generale: non è data infatti una definizione della "qualità della ricerca" che non faccia riferimento ad essi. Ciò nondimeno vi sono altri passaggi della costruzione di questo concetto che possono essere presi in considerazione indipendentemente dalla definizione dei singoli criteri.

Sartori (1984, p. 63-64[54]) a chiusura dei lavori del COCTA (*Committee on Conceptual and Terminological Analysis*), ha proposto una serie di regole volte a da un lato a evitare la proliferazione concettuale, dall'altro a perfezionare la costruzione e la ricostruzione concettuale nelle scienze sociali entro un quadro di argomentabilità, correttezza e adeguatezza. Le regole proposte costituiscono un decalogo:

1: «Di ogni concetto controllare sempre, e separatamente, (1) se è ambiguo, cioè, come il significato si collega con il significante; e (2) se è vago, cioè, come il significato si collega con il referente».

[53] Carnap (1936) ha introdotto la questione delle proposizioni di riduzione in relazione alle proprietà *disposizionali* (presentate dagli oggetti solo a date condizioni) o *quantitative* (in grado di assumere un numero di valori potenzialmente infinito, e dunque non precisamente rilevabili), «il cui significato può essere determinato solo parzialmente e condizionatamente» (Agnoli, 1994, p. 93).

[54] Traduzione dall'originale in lingua inglese.

2a: «Controllare sempre (1) se i termini chiave (il designatore del concetto e i termini collegati) sono definiti; (2) se il significato designato dalla loro definizione è inequivocabile; e (3) se il significato rimane invariato (cioè coerente) per tutta l'argomentazione»;

2b: «Controllare sempre se i termini chiave sono utilizzati in modo univoco e coerente nel significato dichiarato».

3a: «A meno di una prova contraria, nessuna parola dovrebbe essere usata come sinonimo di un'altra parola».

3b: «Per quanto riguarda la stipulazione di sinonimie, l'onere della prova è rovesciato: quello che richiede dimostrazione è che, attribuendo differenti significati a termini differenti, si crei una distinzione priva di conseguenze».

4: «Nel ricostruire un concetto, per prima cosa raccogliere un insieme rappresentativo di definizioni; per seconda, estrarre le loro caratteristiche; e per terza, costruire matrici che organizzino queste caratteristiche in modo significativo».

5: «Con riferimento all'estensione di un concetto, valutare sempre (1) il suo grado di vastità (*boundlessness*), e (2) il suo grado di discriminazione denotativa nei confronti dell'insieme cui si riferisce».

6: «La vastità di un concetto viene ridotta incrementando il numero delle sue proprietà; e la sua capacità di discriminare viene migliorata quando vengono immesse proprietà aggiuntive».

7: «L'estensione e l'intensione di un concetto sono inversamente proporzionali».

8: «Nella scelta del termine che designa il concetto, fare sempre riferimento a un controllo del campo semantico a cui il termine appartiene – cioè all'insieme di parole simili che vi sono associate».

9: «Se il termine che designa il concetto sconvolge il campo semantico (cui il termine appartiene), allora giustificare la scelta mostrando che (1) nessuna porzione di significato viene perduta, e che (2) l'ambiguità non aumenta a causa del trasferimento».

10: «Assicurarsi che il *definiens* di un concetto sia adeguato e parsimonioso: adeguato in quanto contiene sufficienti caratteristiche per identificare i referenti e i loro confini; parsimonioso in quanto nessuna proprietà accessoria è inclusa tra le proprietà definitorie necessarie».

La prima regola richiede di controllare se il concetto sia ambiguo e se sia vago, connotandosi come una sorta di regola generale, ed è possibile assicurarsi il suo rispetto seguendo alcune delle regole successive, in particolare la seconda e la terza in riferimento all'ambiguità, la quinta la sesta e la settima in riferimento alla vaghezza (Fasanella, 1993). Infatti con le prime due «si richiede di controllare che (a) i termini usati siano dotati di significato, (b) il significato dia univoco e costante, (c) all'introduzione di nuovi termini corrisponda effettivamente l'introduzione di nuovi significati» (Fasanella, 1993, p. 171); la quinta, la sesta e la settima «prescrivono di (a) individuare chiaramente i confini del concetto sulla base di un numero sufficientemente ampio di proprietà che definiscono il concetto stesso (il problema della *boundary indefiniteness*), (b) individuare senza difficoltà gli elementi che possono essere trattati in qualità di membri rientranti nell'estensione del concetto (il problema della *membership indefiniteness*), (c) stabilire i valori limite che l'oggetto può assumere su una determinata proprietà che definisce il concetto (il problema della *cut-off indefiniteness;* Sartori, 1984, p. 42)» (*ibidem*). Seguendo ancora Fasanella è possibile però notare che «mentre tali norme regolano chiaramente materie riguardanti la definizione denotativa, precisante e operativa, le rimanenti regole (la numero 4, 8, 9, 10) sono da applicare al campo della costruzione e soprattutto

della ricostruzione concettuale» (*ibidem*, p. 172). Nel caso in analisi dunque queste quattro regole sono da riferire al concetto generale, mentre la riflessione circa l'ambiguità e la vaghezza va riservata alla definizione dei singoli criteri (*cfr*. § 2.2).

La quarta regola del decalogo sartoriano raccomanda nella ricostruzione di un concetto di fare riferimento alle definizioni già disponibili, attraverso alcuni passaggi: «per prima cosa raccogliere un insieme rappresentativo di definizioni; per seconda, estrarre le loro caratteristiche; e per terza, costruire matrici che organizzino queste caratteristiche in modo significativo» (Sartori, 1984, p. 64). Nel caso specifico della qualità della ricerca è importante fare riferimento a definizioni utilizzate in contesti simili: cioè al fine di valutare prodotti scientifici nell'ambito di esercizi di valutazione nazionali. Definizioni utilizzate in relazione alla valutazione di progetti di ricerca o specifici programmi non risulterebbero infatti adeguati agli scopi conoscitivi in relazione ai quali il concetto è stato elaborato.

Gli esercizi di valutazione della ricerca condotti livello nazionale sono numerosi, ma risultano fortemente differenziati da Paese a Paese, non solo per i loro approcci valutativi e obiettivi conoscitivi, ma anche per l'impatto, più o meno rilevante, che i loro risultati possono avere sul finanziamento di strutture e progetti di ricerca. Lo scopo di questa regola sartoriana è l'individuazione delle «caratteristiche fondamentali» del concetto (*ibidem*, p. 42) sulla base di un set di definizioni rappresentative, ma non essendo possibile individuare un insieme di definizioni rappresentative, data l'entità e la quantità di differenze riscontrabili tra i diversi esercizi di valutazione, ci si affiderà qui alla definizione in uso per quello più autorevole e consolidato: l'esercizio britannico.

Nel Regno Unito, dove la valutazione della ricerca è stata implementata già a partire dalla seconda metà degli anni '80[55], l'esercizio di valutazione della ricerca scientifica nazionale è il REF *Research Excellence Framework* che da poco ha sostituito il RAE *Research Assessment Exercise*. Uno dei punti più discussi del RAE, che hanno spinto l'HEFCE a modificare l'impianto dell'esercizio di valutazione, era la mancanza di coerenza nei criteri di valutazione (McNay, 2003), per questa ragione il REF ha previsto fin dal 2011 una consultazione per la definizione dei criteri che ha coinvolto in primo luogo i panel e i sub-panel disciplinari, ma anche le comunità scientifiche e gli "utilizzatori" della ricerca. I risultati di queste consultazioni sono stati tradotti in modifiche ai criteri e alle procedure (REF 01.2012). Questo punto è importante perché rappresenta la messa in pratica di quelle fasi di partecipazione, negoziazione e condivisione che preludono a una ricerca valutativa effettivamente valida (Palumbo, 2001). In Italia per la VQR, almeno con riferimento ai criteri di valutazione generale[56], non vi è stata alcuna consultazione della comunità scientifica, forse a causa

[55] Al primo esercizio di valutazione condotto nel Regno Unito nel 1986, sotto il governo di Margaret Thatcher, hanno fatto seguito due esercizi denominati *Research Selectivity Exercise*, condotti nel 1989 e nel 1992, e tre tornate di RAE *Research Assessment Exercise* (nel 1996, 2001 e 2008) condotto congiuntamente da HEEFCE (*Higher Education Funding Council for England*), SFC (*Scottish Funding Council*), HEFCW (*Higher Education Funding Council for Wales*) e DEL (*Department for Employment and Learning, Northern Ireland*).

[56] Le comunità scientifiche e professionali sono state consultate in riferimento a specifici passaggi, come la classificazione delle riviste nelle Aree non bibliometriche, e quanto almeno in alcune Aree tra cui l'Area 14, nonostante questo confronto preliminare «la polemica non si è completamente sopita» (Anvur, 2013d, GEV 14, p. 29; cfr. Capitolo 4).

della strettezza dei tempi in cui l'esercizio VQR è stato progettato e portato a termine[57]. A proposito dei tempi di realizzazione della VQR infatti nel rapporto finale dell'Anvur si legge: «averla portata a termine in poco più di diciotto mesi ha certamente comportato a volte il sacrificio dell'ottimo a favore del "buono", ma, al di là delle prescrizioni del DM sui tempi di realizzazione, la comunità nazionale dei ricercatori attendeva da molto tempo un'analisi accurata dello stato della ricerca nel nostro paese, e quindi non ci si poteva permettere, come ad esempio è avvenuto nel Regno Unito per il *Research Excellence Framework*, di dedicare anni alla sua preparazione e alla sperimentazione» (Anvur, 2013a, p. 11-12).

Nei documenti di presentazione del REF non è data nessuna definizione di "qualità della ricerca"; come nel caso della VQR infatti la qualità viene definita attraverso l'esplicitazione delle sue dimensioni. Nel REF la qualità complessiva della ricerca è data da:

- la qualità dei suoi output, in termini di originalità, rilevanza e rigore in relazione agli standard di qualità internazionali;
- la qualità dei suoi effetti, in termini di capacità e rilevanza;
- la qualità dell'ambiente in termini di sostenibilità e vitalità (REF 02.2011, p. 6[58]).

E' il primo di questi punti a costituire il termine di confronto per la definizione della qualità della ricerca nella VQR: evidentemente i prodotti corrispondono agli *output* della ricerca. Nel REF la valutazione è affidata a quattro panel (A, B, C, D) ognuno dei quali fornisce una propria definizione denotativa dei criteri, calibrata in base alle specifiche caratteristiche delle discipline ma fondata su una base comune; per brevità e similarità con le Aree non bibliometriche nella VQR, riportiamo qui quella del panel C[59]:

- «L'*originalità* sarà intesa nei termini del carattere innovativo del prodotto della ricerca. Un prodotto della ricerca che dimostri originalità può: confrontarsi con problemi nuovi e/o complessi, sviluppare metodi di ricerca, metodologie e tecniche di analisi innovativi, e/o far avanzare la teoria o l'analisi della dottrina, dei criteri o delle pratiche» (REF 01.2012, p. 66-67[60]).
- «La *rilevanza* sarà intesa in termini di sviluppo dell'agenda intellettuale del campo e può essere teorica, metodologica e/o sostantiva. Il peso dovuto verrà assegnato alla rilevanza potenziale quanto a quella attuale, soprattutto nel caso in cui il prodotto sia molto recente» (*ibidem*).
- «Il *rigore* sarà inteso in termini di precisione intellettuale, robustezza e adeguatezza dei concetti, dell'analisi, delle teorie e delle metodologie implementate all'interno di un prodotto della ricerca. Si terrà conto di qualità come la completezza, la coerenza e la robustezza delle argomentazioni e delle analisi, quanto della dovuta considerazione delle questioni etiche» (*ibidem*).

La definizione del REF risulta meno ambigua e vaga di quella messa a punto dall'Anvur, principalmente grazie a una serie di puntuali specificazioni del significato che contribuiscono a connotare i confini di ciascuna dimensione del concetto e ad identificarne i referenti empirici (a

[57] L'Anvur è stato regolamentato con Decreto Ministeriale il 1 Febbraio 2010. Il bando per la partecipazione alla VQR è uscito il 7 novembre 2011, e il rapporto finale è stato pubblicato il 22 Luglio 2013.
[58] Traduzione dall'originale in lingua inglese.
[59] Questo panel copre le discipline che nella VQR sono incluse nelle Aree 8, 12, 13 e 14, più alcune discipline dell'Area 11 (in particolare le scienze pedagogiche).
[60] Traduzione dall'originale in lingua inglese.

questo proposito si veda il § 2.2). Per grandi linee, riflettendo cioè solo sulle dimensioni della qualità della ricerca considerate ai fini della valutazione, la differenza più evidente è nella mancanza del riferimento al rigore nella definizione italiana (Tabella 12). Vi è un riferimento al rigore nella formulazione delle classi di merito *eccellente* e *buono*, tuttavia in nessuna delle definizioni dei criteri è presente un chiaro riferimento al rigore metodologico o etico.

Tabella 12 – Confronto tra le dimensioni della qualità dei prodotti della ricerca nel REF e nella VQR

REF	VQR
Rilevanza	Rilevanza
Originalità	Originalità
Rigore	-
-	Internazionalizzazione

Una seconda differenza è la mancanza nel REF del riferimento all'internazionalizzazione. E' vero che nella esplicitazione dei criteri la qualità dei prodotti è definita: «in termini di originalità, rilevanza e rigore in relazione agli standard di qualità internazionali» (REF 02.2011, p. 6[61]), tuttavia in questo caso gli standard internazionali sono il termine di paragone su cui fondare la valutazione dell'originalità, della rilevanza e del rigore del prodotto; rappresentano un «posizionamento nello scenario internazionale» (Anvur, 2011, p. 7), ma non nel senso della collocazione o della competitività editoriale. La differenza può sembrare sottile, ma non lo è: nel primo caso il prodotto viene confrontato con gli standard internazionali di originalità, rilevanza e rigore; nel secondo il prodotto viene valutato per la sua internazionalità (in termini di «rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica»; *ibidem*) e non solo per la sua rispondenza agli standard internazionali. La differenza è che nel primo caso si tiene conto della qualità del prodotto in sé, nel secondo si tenta di dedurla dalle sue caratteristiche più (competitività e diffusione editoriale) o meno (rilevanza e apprezzamento della comunità scientifica) manifeste.

Vale la pena di chiedersi qual è il rapporto tra la dimensione dell'internazionalità e la qualità della ricerca, anche se rimandiamo la riflessione sulla relazione tra questa e le altre due dimensioni al prossimo paragrafo. Leggendo le definizioni delle classi di merito si ha l'impressione che l'internazionalità sia una caratteristica dei prodotti che non influenza necessariamente la loro qualità, per tutte le classi si fa infatti riferimento tanto al piano internazionale quanto a quello nazionale (anche nel caso della classe *limitato*, anche se qui il riferimento è a una «sede internazionale di non particolare rilevanza»*;* Anvur, 2011a, p. 7), mentre assume un peso pari agli altri due criteri nel determinare dell'esito della valutazione (*cfr.* Capitolo 4).

Le regole 8 e 9 del decalogo sartoriano risultano legate tra loro, infatti la prima raccomanda di «fare sempre riferimento a un controllo del campo semantico a cui il termine appartiene – cioè all'insieme di parole simili che vi sono associate» (Sartori, 1984, p. 64) nella scelta del termine che designa il concetto, la seconda di porre attenzione ai casi in cui la scelta del termine ne sconvolga il campo semantico, e raccomanda di «giustificare la scelta mostrando che (1) nessuna porzione di significato viene perduta, e che (2) l'ambiguità non aumenta a causa del trasferimento» (*ibidem*). Entrambe queste regole hanno una scarsa influenza sulla costruzione del concetto di qualità della ricerca che sostanzialmente, come già evidenziato, è in se un concetto contestuale e relativo, che

[61] Traduzione dall'originale in lingua inglese.

necessita più che di una definizione vera e propria, di una riduzione che aiuti a individuarne i referenti. Non si tratta di selezionare il termine da associare al significato, ma viceversa di limitare i significati connessi a quel termine in un contesto specifico (la VQR).

L'ultima regola invece risulta estremamente rilevante, e prescrive di assicurarsi «che il *definiens* di un concetto sia adeguato e parsimonioso: adeguato in quanto contiene sufficienti caratteristiche per identificare i referenti e i loro confini; parsimonioso in quanto nessuna proprietà accessoria è inclusa tra le proprietà definitorie necessarie» (Sartori, 1984, p. 64). Con riferimento a quanto considerato circa il criterio dell'internazionalizzazione è possibile osservare che il concetto di qualità della ricerca nella VQR non risulta parsimonioso proprio perché questo criterio si configura come una proprietà accessoria e non come una proprietà definitoria necessaria. Inoltre il *definiens* in questione non è completamente adeguato, analizzando la formulazione dei singoli criteri (*cfr*. § 2.2) risulterà infatti evidente che tanto l'ambiguità quanto la vaghezza delle definizioni non permettono una identificazione univoca dei referenti empirici cui le proprietà in questione sono riferibili.

3.2 Le dimensioni concettuali della qualità della ricerca

L'individuazione delle dimensioni del concetto è il passo che, nel modello lazarsfeldiano per l'operativizzazione[62], segue la rappresentazione figurata del concetto, tuttavia nel caso della qualità della ricerca i due passaggi sono in realtà simultanei. E' stato già evidenziato che per costruire un concetto relativo e contestuale come quello di "qualità" è necessario individuare (o decidere) i criteri che lo definiscono: le dimensioni in questo caso *sono* il concetto. In questo caso, come in altri nella ricerca sociale, «il concetto costituisce il *definiendum*, e il prodotto logico degli aspetti o dimensioni costitutive rappresenta il *definiens*» (Fasanella, 1993, p. 174).

La prima questione da affrontare è quella delle relazioni semantiche: non solo quella che si instaura tra le diverse dimensioni e il concetto, ma anche quelle che sussistono tra una dimensione e l'altra. Si tratta di controllare che ciascuna delle dimensioni rispetti i requisiti previsti dalla prima regola di Sartori: «di ogni concetto controllare sempre, e separatamente, (1) se è ambiguo, cioè, come il significato si collega con il significante; e (2) se è vago, cioè, come il significato si collega con il referente» (1984, p. 63), tenendo conto anche delle relazioni reciproche e di quelle con il *definiens*.

Le definizioni dei singoli criteri, nonostante non stipulino sinonimie (regola 3, Sartori, 1984, p. 63), non risultano del tutto esenti da ambiguità, è infatti possibile rinvenire alcuni indizi che indicano una incompleta definizione dei significati attribuiti ai termini, e una non univocità della relazione tra termini e significati.

La definizione dei termini risulta incompleta, in quanto non tutti i termini chiave sono definiti: ad esempio definendo la rilevanza si fa riferimento a «congruità, efficacia, tempestività e durata delle ricadute», ma nessuna di queste caratteristiche viene ulteriormente precisata. Le ambiguità che ne derivano possono essere ricondotte all'alveo delle problematiche legate all'omonimia (regola 2, Sartori, 1984, p. 63), infatti in mancanza di una definizione ciascuno dei termini può assumere significati diversi (ad esempio la *tempestività* può essere intesa come *tempismo* o come *celerità*). Inoltre le definizioni non sempre contengono riferimenti puntuali al

[62] Il riferimento è alla procedura in quattro fasi, nota come *paradigma di Lazarsfeld*, che va dall'immagine figurata del concetto alla costruzione dell'indice sintetico (Lazarsfeld, 1969).

contesto di riferimento dei termini non permettendone una interpretazione univoca (ad esempio le *ricadute* possono essere intese come socio-economiche, teoriche, applicative, e così via).

Si presenta anche un problema di sinonimia, particolarmente rilevante dato che deriva dal fatto che la definizione del criterio di *rilevanza* e quella del criterio di *originalità/innovazione* risultano parzialmente sovrapponibili:

a) *rilevanza*: «da intendersi come **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» (Anvur, 2011, p. 7);
b) *originalità/innovazione*: «da intendersi come **contributo all'avanzamento di conoscenze** o a nuove acquisizioni nel settore di riferimento» (*ibidem*).

Questa parziale sovrapponibilità indica che uno stesso significato è associato a due termini diversi, nello specifico che il significato attribuito al termine *originalità/innovazione* è ricompreso nel significato del termine *rilevanza*.

Nella definizione dell'internazionalizzazione inoltre si fa riferimento al «posizionamento nello scenario internazionale, in termini di **rilevanza**, competitività, diffusione editoriale e apprezzamento della comunità scientifica» (*ibidem*). Qui è possibile ipotizzare due diverse situazioni: o il termine *rilevanza* viene utilizzato con un significato differente da quello definito e siamo di fronte a un problema di omonimia, oppure il termine è utilizzato con lo stesso significato e in questo caso il criterio della *rilevanza* copre anche parte del campo semantico del criterio dell'*internazionalizzazione*.

Si è detto che l'ambiguità è riferibile alla relazione instaurata tra termine e significato, mentre la vaghezza è riferibile alla relazione tra il concetto e i suoi referenti empirici, ma queste due caratteristiche non sono slegate: tanto più infatti un concetto appare ambiguo tanto meno risulteranno circoscrivibili i suoi referenti empirici.

Sempre facendo riferimento a Sartori (1984), considerando l'intensione e l'estensione dei criteri è possibile notare che i problemi evidenziati circa la sovrapposizione del significato del termine rilevanza agli altri due criteri conducono a una *boundary indefiniteness*, cioè a una mancanza di chiarezza nella definizione dei confini del concetto. Da un lato, infatti, criterio per criterio le proprietà del concetto indicate non sono sufficienti a delimitarlo, dall'altro queste proprietà vengono riferite contemporaneamente a più criteri, sfocando i confini tra l'uno e l'altro.

Oltre a questa mancanza di demarcazione concettuale anche la scarsità di riferimenti puntuali e la sotto-determinazione del significato attribuibile alle definizioni rendono poco immediata l'individuazione degli elementi che possono essere considerati come parte dell'estensione del concetto, creando un problema di *membership indefiniteness*.

Tenendo presente però il fine valutativo del costrutto in questione il problema più rilevante è indubbiamente quello relativo alla *cut-off indefiniteness*, vale a dire della vaghezza circa *quanto* un oggetto debba possedere di ciascuna delle proprietà del concetto per essere ricompreso tra i sui referenti. Ad esempio: quanto un prodotto deve contribuire all'avanzamento di conoscenze per poter essere definito originale? Chiaramente possiamo immaginare che tutti i prodotti abbiano un certo grado di rilevanza, originalità e internazionalizzazione, infatti l'Anvur, oltre a fornire una definizione dei criteri, ha fornito anche una definizione delle classi di merito (*cfr.* § 1.1; Anvur, 2011a, p. 7). Queste definizioni tuttavia non risolvono il problema relativo alla vaghezza, se infatti è vero che porgono dei riferimenti è vero anche che questi riferimenti risultano poco chiari: per i prodotti di

livello limitato si parla di «contributo modesto», per quelli accettabili di «accresciuto di qualche misura», per quelli buoni il riferimento è all'«importanza» e per quelli eccellenti all'«eccellenza» e al rinnovo «in maniera significativa» delle conoscenze nel campo (*ibidem*). Inoltre come nel caso del REF viene esplicitato il riferimento agli standard di qualità internazionali per stabilire un termine di confronto, nel caso della VQR il riferimento è alla «scala di valore condivisa a livello internazionale» (*ibidem*). In parte questa vaghezza è comprensibile, dato che la formulazione dei criteri deve servire da guida a soggetti che possiedono già una cognizione di ciascuno di essi e che sono selezionati proprio in quanto pari, cioè in quanto partecipi di una comunità che condivide una serie di valori e regole (Fasanella, 2013), ciò nonostante può riflettersi nella precisione della definizione operativa (*cfr.* Capitolo 4).

I problemi di ambiguità e vaghezza appena evidenziati sono di una certa rilevanza non solo dal punto di vista prettamente semantico, ma anche con riferimento alla questione delle relazioni logiche e pragmatiche tra le dimensioni. Le relazioni tra le dimensioni non sono infatti esclusivamente di natura semantica, ma anche di natura logica e pragmatica (Statera, 1997).

Dal punto di vista logico è infatti possibile individuare tra le dimensioni concettuali nessi di congiunzione, disgiunzione, negazione, implicazione o co-implicazione che definiscono la struttura del concetto a cui le dimensioni sono riferite. Vale la pena di riportare nuovamente la definizione delle classi di merito fornita nel bando della VQR, per poter presentare direttamente alcune osservazioni:

«1. I prodotti di livello *eccellente* sono quelli riconosciute 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.
2. I prodotti di livello *buono* sono quelli di importanza internazionale e nazionale riconosciute per originalità dei risultati e rigore metodologico.
3. 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.
4. 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» (Anvur, 2011a, p. 7).

Il legame tra le dimensioni della qualità non risulta esposto chiaramente, così come le relazioni tra le dimensioni. Innanzitutto nella formulazione degli asserti non è sempre identificabile il riferimento a tutti e tre i criteri, questo per due ragioni: la parziale sovrapponibilità tra la definizione della *rilevanza* e quella dell'*originalità* e la mancata rispondenza tra i termini utilizzati per la formulazione dei criteri e quelli utilizzati per la definizione delle classi di merito.

Circa il primo punto, ad esempio, è possibile notare che definendo il livello *accettabile* si fa riferimento all'accrescimento del patrimonio delle conoscenze, legato tanto alla dimensione della *rilevanza* quanto a quella dell'*originalità*, ma non viene esplicitato se questo riferimento rimandi a entrambi i criteri o a uno solo dei due. La stessa questione si pone anche per la definizione del livello *limitato*.

Circa il secondo punto invece si pensi alla definizione del livello *buono*: è rinvenibile chiaramente un riferimento all'originalità, ma la rilevanza potrebbe essere invece rinvenibile tanto nel riferimento all'*importanza* quanto in quello al *rigore metodologico*, che tuttavia non rimandano a nessuno dei tre criteri, o almeno non alla definizione che ne viene fornita nel bando VQR.

Le relazioni tra le dimensioni risultano vaghe proprio perché le loro definizioni sono ambigue: in particolare la mancanza di una demarcazione tra il concetto di *rilevanza* e quello di *originalità* produce nella lettura delle definizioni delle classi di merito l'impressione che tra queste due dimensioni sussista un rapporto di co-implicazione. Non solo cioè la struttura del concetto sembra indicare che se un prodotto sia di qualità solo se risulta sia rilevante sia originale, ma che un prodotto che sia rilevante deve essere anche originale, e che un prodotto originale deve essere anche rilevante. Stando alle definizioni delle classi di merito, in altre parole, le dimensioni della qualità non sono assunte come ortogonali, ma come interdipendenti.

Questa impressione deriva (o dipende) anche dal fatto che nella formulazione delle classi di merito non vengono completamente rispettati i requisiti logici della classificazione: esaustività, mutua esclusività e unicità del *fundamentum divisionis* (Marradi, 1990a).

La definizione delle classi non è esaustiva dato che, ad esempio, non considera il caso di prodotti rigorosi da un punto di vista metodologico ma non originali, oppure quello di prodotti originali ma non rilevanti. Le classi, proprio in quanto la loro definizione non è esaustiva, non risultano neppure mutuamente esclusive: un prodotto potrebbe ricadere nella classe *eccellente* quanto a rilevanza e nella classe *accettabile* quanto a originalità. Entrambi questi problemi derivano dalla mancanza di un unico *fundamentum divisionis*; essendo la classificazione basata su tre criteri, perché il *fundamentum* risultasse unico, sarebbe stato opportuno costruire uno spazio di attributi e assegnare a ciascuno dei tipi da questo risultante una delle classi (procedendo o meno a una loro riduzione funzionale. In Tabella 13 riportiamo uno spazio di attributi esemplificativo, in cui ciascun criterio è rappresentato dicotomicamente con la rispondenza (+) non rispondenza (-), allo scopo di dare un'idea di quante possano essere le combinazioni degli stati dei prodotti sulle proprietà.

Tabella 13 – Spazio degli attributi: rilevanza, originalità e internazionalizzazione (presenza/assenza)

Rilevanza	*Originalità*	*Internazionalizzazione*	
		+	-
+	+	(1)	(2)
	-	(3)	(4)
-	+	(5)	(6)
	-	(7)	(8)

In realtà i tipi teorici possibili sono molto più numerosi, dato che almeno i primi due criteri sembrerebbe graduabili come: «modesto», «di qualche misura», «importante», «eccellente» o «significativo» (Anvur, 2011a, p. 7); anche graduando in quattro classi ciascun criterio, solo in riferimento a una piccola parte di questi tipi viene definita una classe di merito, come risulta evidente dalla ricostruzione grafica in Tabella 14.

Il ruolo assunto dall'internazionalizzazione nella determinazione del livello di qualità va approfondito: nella rappresentazione grafica, solo nel caso della classe limitato questo criterio assume una funzione classificatoria. Stando alla definizione delle singole classi di merito questo criterio risulterebbe infatti sostanzialmente ininfluente: i prodotti di livello *eccellente* sono infatti «riconosciuti come eccellenti a livello **internazionale** […] oppure quelli che hanno rinnovato in maniera significativa il campo degli studi a livello **nazionale**» (Anvur, 2011a, p. 7), quelli di livello *buono* sono «di importanza **internazionale e nazionale**» (*ibidem*), quelli di livello *accettabile* sono «a diffusione **internazionale o nazionale**» (*ibidem*), quelli di livello *limitato* sono «diffusione **nazionale o locale**, oppure in sede **internazionale** di non particolare rilevanza» (*ibidem*). Si noti che è la

specificazione di *non particolare rilevanza* a determinare il ruolo classificatorio del criterio nella Tabella 14, tuttavia se si considerasse l'internazionalizzazione come una proprietà dicotomica, o comunque se ne fornisse una definizione operativa non in grado di rilevare il grado di rilevanza della sede di pubblicazione, questa dimensione concettuale risulterebbe a tutti gli effetti ininfluente ai fini della determinazione della classe di merito.

Tabella 14 – Spazio degli attributi: rilevanza, originalità e internazionalizzazione (con 4 gradienti)[63]

Rilevanza	*Originalità*	*Internazionalizzazione*			
		++	+	-	--
++	++	E	E	E	E
	+				
	-				
	--				
+	++				
	+	B	B	B	B
	-				
	--				
-	++				
	+				
	-	A	A	A	A
	--				
--	++				
	+				
	-				
	--		L	L	L

La relazione tra le dimensioni e il concetto generale e le relazioni tra le dimensioni non sono dunque stabilite con chiarezza, e questa questione assume una importanza centrale non solo dal punto di vista concettuale, ma anche dal punto di vista operativo. Infatti il passaggio relativo alla selezione delle dimensioni concettuali «anticipa direttamente il successivo, concernente l'individuazione e la scelta degli indicatori, di fatto lo implica» (Agnoli, 1994, p. 121). Per questa ragione l'individuazione delle dimensioni è anche un processo pragmatico: nell'operarla è necessario tenere conto non solo degli obiettivi della ricerca e dello schema teorico di riferimento ma anche delle già citate «valutazioni di maggiore o minore, più precisa o meno precisa, traducibilità operativa» (Statera, 1997, p. 129).

Conclusioni

La definizione del concetto di qualità della ricerca e della sue dimensioni rilevanti ai fini della valutazione dei prodotti nel corso della VQR presenta eccessivi margini di ambiguità e vaghezza. La parziale sovrapponibilità delle definizioni di rilevanza e originalità, la mancanza di corrispondenza tra la definizione dei criteri e quella delle classi di merito, il mancato rispetto dei requisiti logici della classificazione nella definizione semantica delle classi, conducono a una serie di criticità: *cut-off indefiniteness*, *membership indefiniteness* e *boundary indefiniteness*.

[63] Qui e in tutte le tabelle che seguono le iniziali indicano le classi di merito corrispondenti: E=Eccellente, B=Buono; A=Accettabile e L=Limitato.

La scelta delle dimensioni va tuttavia considerata anche e soprattutto alla luce dello scopo dell'indagine: la valutazione delle strutture dedicate alla ricerca scientifica in Italia. La ricerca valutativa presenta infatti delle peculiarità che devono necessariamente essere considerate nell'analisi dei concetti che vi sono impiegati. I requisiti della valutazione sono: « la produzione di un giudizio fondato sull'intenzionalità dell'azione da valutare (o sulla ricostruibilità della razionalità, strumentale o valoriale, di tale azione o insieme di azioni) e la disponibilità di criteri di giudizio, nonché il fatto che l'azione realizzata permetta la raccolta di riscontri empirici utili a supportare il giudizio stesso» (2001, p. 48). La VQR prevede la valutazione esterna di azioni intenzionali realizzate, sulla base di una serie di criteri; in questa sede non verrà discussa la questione dell'intenzionalità/razionalità della ricerca, e si darà per scontato che la ricerca scientifica produca una serie di riscontri che possono essere utilizzati come una base empirica per la formulazione di giudizi; ciò che qui interessa maggiormente sono naturalmente i criteri di giudizio.

Dato che non si intende discutere nel merito la scelta dei criteri, la questione centrale è relativa alla traducibilità operativa delle singole dimensioni. Questa dipende direttamente dalla chiarezza delle definizioni e allo stesso tempo determina diverse caratteristiche delle loro definizioni operative. Tanto più una definizione è ambigua tanto meno sarà traducibile in termini empirici, tanto più risulterà vaga, tanto meno precisa sarà la sua traduzione.

Sarà dato ampio spazio alle definizioni operative di qualità della ricerca utilizzate nella VQR, ma circa le conseguenze pragmatiche della selezione delle dimensioni vale la pena ricordare che la valutazione deve essere fondata su procedure rigorose e codificabili (Palumbo, 2001). Se i criteri presentano definizioni lessicali vaghe e/o ambigue, e dunque non sono completamente o precisamente traducibili in definizioni operative, il grado di rigore procedurale della valutazione risulta ridotto, e i giudizi formulati risultano meno legati all'esito di un procedimento scientifico e più vulnerabili all'influenza di precognizioni e valori soggettivi.

Capitolo 4
La definizione operativa della qualità della ricerca

Introduzione

Il capitolo presenta l'analisi metodologica dell'operativizzazione, strettamente intesa, del concetto di qualità della ricerca nella VQR. Sono dunque approfondite le questioni relative alla traduzione delle dimensioni concettuali prima in indicatori e poi in variabili, e quelle legate alla ricomposizione delle informazioni rilevate in un indice sintetico. In riferimento alla valutazione dei prodotti nella VQR quest'ultimo passaggio corrisponde all'assegnazione di ciascun prodotto a una "classe di merito" (*Eccellente*, *Buono*, *Accettabile*, *Limitato* oppure *Non valutabile*), ma come si è già evidenziato le procedure che lo precedono sono differenti da Area a Area, e a volte anche tra diversi aettori scientifico-disciplinari all'interno di una stessa Area. In questa fase si far, dunque, riferimento soprattutto ai due casi studio: l'Area 14 delle Scienze Politiche e Sociali e l'Area 3 delle Scienze Chimiche.

Nel corso della trattazione sono esposte e approfondite le principali questioni metodologiche connesse a ciascuna delle fasi del processo di operativizzazione, al fine di introdurre l'analisi delle procedure poste in atto nel corso della VQR per ciascuna Area selezionata come caso di studio.

Le definizioni operative, come i concetti, «non sono né vere né false» (Marradi, 1980, p. 25), tuttavia «ogni ricercatore è perfettamente libero di dare la definizione operativa che vuole di un determinato concetto di proprietà, purché sottoponga tale sua scelta, quando pubblica i risultati di una ricerca, al giudizio della comunità scientifica» (*ibidem*). La trasparenza delle procedure è un nodo cruciale in riferimento a questo passaggio, ciò nonostante nel caso della VQR diversi passaggi sono scarsamente argomentati o rendicontati.

4.1 Dalle dimensioni concettuali agli indicatori

Marradi (2007) riconduce l'uso di indicatori nelle scienze sociali alla necessità di non rinunciare a raccogliere informazioni su una certa proprietà nel caso in cui non sia possibile immaginare una definizione operativa diretta, oppure nel caso in cui una definizione operativa diretta non risulterebbe plausibilmente affidabile. Un indicatore dunque deve ammettere accettabili definizioni operative dirette e presentare una stretta relazione semantica con la proprietà che interessa rilevare, cioè deve essere possibile stabilire un *rapporto di indicazione* con il concetto. La centralità del rapporto di indicazione non dipende esclusivamente dalla maggiore o minore distanza tra concetto e referente empirico: questo problema si sviluppa in ragione della complessità e della multidimensionalità degli oggetti d'indagine, oltre che dalla latenza delle grandezze studiate (Cannavò, 1999). Oggetti e concetti complessi e multidimensionali pongono sfide teoriche prima

ancora che metodologiche, connesse innanzitutto alla concettualizzazione e solo in secondo luogo al problema della traduzione in termini empirici.

Ogni concetto che non suggerisce direttamente una definizione operativa richiede una pluralità di indicatori (pluralità verso il basso) e allo stesso tempo ogni concetto che può essere direttamente operativizzato può essere scelto come indicatore di una pluralità di altri concetti (pluralità verso l'alto; Marradi, 2007). La pluralità verso l'alto implica che un unico concetto direttamente operativizzabile sia legato a più concetti non direttamente operativizzabili e che dunque solo parte del suo significato sia riconducibile a quello che si intende effettivamente rilevare. La parte indicante di un concetto direttamente rilevabile è quella realmente connessa con il concetto che si intende rilevare, ciò che invece può risultare legato ad altri concetti è definibile come parte estranea; perché un rapporto di indicazione sia valido la sua parte indicante deve essere il più estesa possibile, quella estranea il più ridotta possibile.

Nella sistematizzazione di Lazarsfeld delle fasi attraverso cui è possibile giungere da un concetto teorico a un indice empirico[64], la fase più discussa è la selezione degli indicatori destinati a rappresentare empiricamente le dimensioni del concetto, scelte in base alla sua definizione. La concezione di Marradi ha in comune con quella di Lazarsfeld la caratterizzazione degli indicatori come osservabili o rilevabili in relazione con un concetto, e anche il riconoscimento del fatto che un indicatore non può essere ritenuto completamente rappresentativo (nei termini di Lazarsfeld, che parla di *campionamento degli indicatori*; 1958) del concetto da operativizzare. Tuttavia mentre per il primo questa relazione è esclusivamente di natura semantica per il secondo è una relazione sostanzialmente probabilistica, sintattica[65]. La definizione di Cannavò (1995, 1999) sembra in grado di mediare tra l'interpretazione marradiana e quella lazarfeldiana del rapporto di indicazione, riferendosi a «un modello semplice di una dimensione concettuale, estendibile solo in via probabilistica al concetto […] in ragione del grado di saturazione semantica della dimensione rispetto al concetto e dell'indicatore rispetto alla dimensione» (1995, pp. 13-14).

Il processo di operativizzazione assume necessariamente la struttura a diamante, simmetrica, che generalmente accompagna la presentazione del modello procedurale lazarsfeldiano: le dimensioni possono essere, in alcuni casi, scomposte in sotto-dimensioni, e la *copertura semantica* di

[64] Si tratta forse del contributo di Lazarsfeld che ha maggiormente inciso sulla pratica della ricerca sociale. Non mancano le critiche a questo protocollo, generalmente indicato come il *paradigma di Lazarsfeld*, legate alla mancanza di qualsiasi riferimento all'origine dei concetti o alla necessità di riferirsi a più livelli di astrazione (tra gli altri Cannavò, 1999).

[65] In *Dagli indicatori agli indici empirici* (1969) Lazasfeld scrive infatti che la relazione tra ogni indicatore e il concetto fondamentale è definita in termini di probabilità, e che è possibile effettuare studi di verifica in relazione alla misura in cui un indicatore è riferibile a un concetto e non a un altro, oppure utilizzarne un certo numero per compensare gli effetti dovuti ai fattori che influenzano gli indicatori ma non sono legati al concetto. Secondo Ricolfi (1992), mentre nell'ottica di Marradi concetti e indicatori sono entrambe entità teoriche, per quanto a differenti distanze dai referenti empirici, dunque il loro rapporto non può che essere semantico; per Lazarsfeld entrambi sono pensabili anche come variabili, dunque il loro rapporto può essere anche di natura sintattica. Ricolfi evidenzia che la scelta tra l'una e l'altra interpretazione dipende in ultima analisi dal giudizio del ricercatore sull'effettivo funzionamento di un certo fenomeno: si procede con controlli standardizzati se si pensa che le relazioni tra gli indicatori non possano essere attribuite a una fonte di variazione comune, in modo standard se pensiamo che queste relazioni possano essere riprodotte in modo adeguato da un modello di attribuzione identificabile (Ricolfi, 1992). Apparentemente, tuttavia, in questo modo si perde di vista il rapporto di indicazione come legame con un concetto per focalizzare invece il rapporto tra indicatori e indici sintetici.

alcune di esse può richiedere più indicatori di altre (Cannavò, 1999). Lo stesso Cannavò introduce una distinzione tra una semantica derivativa o causale degli indicatori (che prevede che il significato del concetto sia saturabile da una somma di indicatori ortogonali tra loro) e una semantica costruttiva o probabilistica (che prevede che il significato del concetto non sia saturabile e indicatori non ortogonali) e dunque la possibilità di una ridondanza informativa (1999)[66]. Pur sostenendo fortemente una concezione della validità come una proprietà del rapporto di indicazione, dunque un rapporto di rappresentatività semantica, Marradi (1980) individua nella coerenza tra indicatori di una stessa dimensione concettuale (o al limite concetto) un "indizio" sia di validità che di attendibilità. Infatti perché un set di indicatori risulti congruente devono verificarsi due condizioni: (a) gli indicatori devono essere in un rapporto di rappresentatività semantica con uno stesso concetto e (b) devono essere stati rilevati in modo attendibile[67].

A proposito del legame tra attendibilità, validità e congruenza di un gruppo si indicatori è di estremo interesse la visione di Campbell e Fiske (1959)[68] il cui vantaggio, tanto rilevante da indurre Marradi a ritenerla estremamente feconda, è nella proposta di affiancare alla validità convergente la validità discriminante, maggiormente legata alla validità per costrutto (perché immediatamente connessa a una riflessione sul contenuto semantico degli indicatori) e in grado di evidenziare (nel caso in cui i tratti siano selezionati adeguatamente in base all'assunto di indipendenza) l'effetto delle tecniche sul risultato della rilevazione.

Un'altra concezione della validità ritenuta convincente da Marradi è la validità secondo costrutto, che assume la corrispondenza tra i risultati ottenuti e una o più ipotesi teoriche come parametro di stima della validità dell'indicatore. L'utilizzo di questa prospettiva necessita di una teoria, o per lo meno di una "generalizzazione empirica" *a là* Merton, sufficientemente specificata da impedire il ricorso a ipotesi *ad hoc* per salvare l'indicatore, e corroborata, in grado cioè di non essere messa in discussione al posto della validità dell'indicatore. In effetti, come sottolineano Fasanella e Allegra (1995), una teoria per giungere a un simile livello di specificazione e consolidamento fa

[66] Se l'interpretazione di Marradi è senza problemi riconducibile a questa seconda concezione, lo stesso Cannavò trova non semplice individuare la posizione di Lazarsfeld, a suo parere ambivalente. I riferimenti all'universo degli indicatori, alla specificazione del significato del concetto sembrerebbero riconducibili ad una semantica derivativa o causale, mentre la concezione probabilistica del rapporto di indicazione e dell'irriducibilità dei termini teorici ai termini empirici implicherebbero una semantica costruttiva o probabilistica.

[67] Una interpretazione del tutto compatibile con la concezione di Lombardo (1994), che afferma che le due questioni sono riferibili a due differenti momenti del farsi della ricerca: la validità è legata alla fase in cui, una volta concettualizzato il problema e le sue dimensioni si stipula il legame tra le dimensioni del concetto e i concetti-indicatori, mentre l'attendibilità è riferibile essenzialmente alle fasi che vanno dalla raccolta dei dati al loro trattamento.

[68] Ogni indicatore è dato dall'unione di una definizione operativa (tecnica) e un certo contenuto semantico (tratto), dunque è possibile che la congruenza di un gruppo di indicatori dipenda non dal contenuto semantico, ma dal fatto che hanno una stessa definizione operativa. Selezionando tecniche e tratti indipendenti e ortogonali, cioè il più dissimili possibili tra loro, è possibile secondo Campbell e Fiske controllare la validità convergente osservando i coefficienti di correlazione di misure differenti dello stesso tratto e la validità divergente osservando i coefficienti di correlazione della stessa misura di tratti differenti. Nei termini di Marradi si avrebbe un indizio di validità, e allo stesso tempo di attendibilità, nel caso in cui indicatori dello stesso concetto registrati con differenti definizioni operative fossero più correlati tra loro di indicatori di concetti diversi registrati tramite la stessa definizione operativa.

necessariamente riferimento a concetti-indicatori già validati, e non può prescindere dalla valutazione del contenuto dei concetti-indicatori.

Meno convincente risulta, sempre in un'ottica marradiana, la validità secondo criterio che presupporrebbe l'esistenza di indicatori già validati. Questo genere di valutazione della validità può configurarsi come predittiva, concorrente, o fare riferimento a gruppi noti, ma in tutti questi casi il prerequisito necessario è la validità dell'indicatore-criterio.

Naturalmente la validità di contenuto, riferibile alla misura in cui un indicatore copre il dominio di significato del concetto che è volto a rilevare, è quella in assoluto più rispondente all'interpretazione di Marradi (1980; 1990a; 2007), e come sostenuto da Fasanella e Allegra (1995), rappresenta un prerequisito necessario all'utilizzo di un indicatore o uno strumento che informa anche tutti i successivi controlli tanto della sua validità quanto della sua attendibilità.

Il concetto di attendibilità, classicamente riferito alla stabilità dello strumento di rilevazione, viene esteso da Marradi (1990b) che, proponendo il termine fedeltà, ne fa una proprietà del rapporto tra il concetto che ha suggerito la definizione operativa e gli esiti effettivi delle operazioni che questa definizione operativa prevede. L'attendibilità/affidabilità può essere riferita ad ogni singolo atto di rilevazione, dunque un indicatore può essere affidabile su un certo numero di casi e non esserlo su altri. In quest'ottica anche un indicatore in stretta relazione semantica con il concetto da rilevare, dunque valido, risulta inutilizzabile se non può essere definito operativamente in modo attendibile.

4.1.1 Gli indicatori di qualità della ricerca nella VQR

L'esplicitazione anche parziale, delle scelte alla base della selezione degli indicatori è un passo cruciale nella rendicontazione di un'indagine, in ragione sia della necessità di pubblicità, riproducibilità e controllabilità del sapere scientifico, sia della necessità di *accountability* connaturata in qualsiasi indagine con fini valutativi.

Campelli, in un più ampio discorso sul metodo, ha espresso incisivamente questo punto: «il metodo comprende in se costitutivamente, passaggi in cui la pubblica razionalità si opacizza o si attenua, connessioni in cui si manifesta una sospensione, si potrebbe dire, di quella integrale dicibilità pubblica che ne costituisce l'immagine ideale […] il discorso scientifico è per così dire un gioco incerto, ma nel quale non è comunque lecito barare. Il fatto che non ogni passo è decidibile in termini rigorosamente razionali, in altri termini, non esime dai requisiti fondamentali di pubblicità e ripercorribilità delle inferenze» (Campelli, 1999, pp. 20-21).

La procedura utilizzata nel corso della VQR per la valutazione dei prodotti presenta diversi punti d'ombra proprio nella pubblicità, riproducibilità e controllabilità della definizione operativa della qualità della ricerca. Circa la scheda di valutazione e gli indicatori utilizzati i rapporti dell'Anvur risultano poco espliciti, e spesso si limitano a un rimando alle appendici che quasi mai però contribuiscono significativamente a chiarire la procedura.

In riferimento alla valutazione in peer review, il rapporto finale prevede tra gli elementi comuni a tutti i GEV: «lo svolgimento guidato della peer review tramite la predisposizione di una scheda di revisione che prevedeva tre domande a risposta multipla pesata» (Anvur, 2013a, p. 22). Grazie a un documento successivo alla pubblicazione dei rapporti finali (Anvur, 2014), si apprende

che la maggior parte delle Aree (2, 3, 6, 7, 8, 10, 11, 12 e 13) ha utilizzato una scala di valutazione con punteggi da 1 a 9, mentre una minoranza (le Aree 1, 4, 5, 9 e 14) ha utilizzato una scala con punteggi da 0 a 3.

Nel rapporto finale, in una nota, si rimanda ai rapporti finali di Area «per le domande e i punteggi» (*ibidem*, p. 26). Tuttavia nella maggior parte dei rapporti di Area ci si limita all'esplicitazione dell'utilizzo di una scheda di valutazione, ad esempio nel rapporto dell'Area 1, Scienze Matematiche e Informatiche, si legge: «i revisori *peer* hanno effettuato la valutazione tramite una apposita scheda-revisore predisposta dal GEV, in cui si chiedeva di valutare la rilevanza, l'originalità e l'impatto del prodotto» (Anvur 2013d, GEV 1, p. 27)[69].

Nel caso dell'Area 8, Ingegneria civile e architettura, «la scheda predisposta in italiano e in inglese chiedeva di assegnare un punteggio compreso tra 1 e 9 in risposta a tre domande, e sollecitava anche l'espressione di un sintetico giudizio scritto» (Anvur 2013d, GEV 8, p. 17). In riferimento all'Area 14, Scienze politiche e sociali, le informazioni fornite sono diverse: «la valutazione dei revisori *peer*, come già detto in precedenza, è basata su un'apposita scheda, predisposta dal GEV, costituita da una serie di domande a risposta multipla e da un campo "commento" facoltativo, mediante la quale i revisori hanno valutato i prodotti sulla base dei criteri di "rilevanza", "originalità/innovazione" e "internazionalizzazione"» (Anvur 2013d, GEV 14, p. 19). Dai rapporti di queste due Aree è possibile almeno individuare alcune caratteristiche delle schede: le domande sono chiuse, a risposta multipla, ad eccezione di un campo facoltativo, dunque le schede presentano un grado di strutturazione elevato.

Nel report dell'Area 11, Scienze storiche, filosofiche, psicologiche e pedagogiche viene precisato che la "scheda valutazione prodotti" è stata predisposta sulla base delle indicazioni Anvur e poi in parte modificata dal GEV, si dichiara che la scheda è riportata nell'Appendice 4 (Anvur 2013d, GEV 11, p. 38), ma in appendice in effetti viene riportato esclusivamente il documento generale dei criteri del GEV (Anvur 2013d, GEV 11, Appendice), dove sono individuabili pressoché le stesse informazioni disponibili per l'Area 14.

Infine nel caso dell'Area 13, Scienze Economiche e Statistiche, e dell'Area 7, Scienze Agrarie e Veterinarie, la scheda viene riportata interamente nelle appendici dei rapporti[70]. Le due schede presentano alcune differenze, ma in entrambi i casi viene posta una sola domanda per ciascun criterio, in forma chiusa, e le modalità di risposta prevedono la graduazione del punteggio assegnato a ciascun prodotto sul singolo criterio da 1 a 9.

La scheda proposta ai revisori di Area 13 è introdotta da una breve premessa, che descrive così il suo contenuto: «per ognuno dei tre criteri (rilevanza, originalità/innovatività, rilevanza/impatto internazionale) viene proposta una lista non esaustiva di domande per chiarirne il

[69] Risultano molto simili a queste le informazioni disponibili sulle schede di revisione per le Aree: 2 Scienze fisiche, 3 Scienze chimiche, 4 Scienze della Terra, 5 Scienze biologiche, 6 Scienze Mediche, 9 Ingegneria industriale e dell'informazione, 10 Scienze dell'antichità, filologico-letterarie e storico-artistiche.

[70] Per l'Area 13 il rapporto contiene sia la scheda (nella versione in inglese) che l'assegnazione delle classi di merito per i punteggi, nell'Appendice E (Anvur, 2013d, GEV 13, Appendice E, p. 112-113), mentre per l'Area 7, sempre in appendice, viene riportata la scheda (anche in questo caso nella versione in inglese), ma non le classi di merito corrispondenti ai punteggi (Anvur, 2013d, GEV 7, Appendice p. 28). In questa sede sarà considerata la forma delle domande e delle modalità di risposta riportate nelle schede, mentre la questione relativa all'assegnazione delle classi di merito sarà approfondita nel § 4.3.

significato[71]» (Anvur, GEV 13, Appendice E, p. 112). A seguito di questo insieme di domande mirate a chiarire il significato del corrispondente criterio viene chiesto al revisore di graduare il risultato della ricerca in termini di *rilevanza* [o *originalità/innovatività*, o *rilevanza/impatto internazionale*] esprimendo un punteggio tra 1 e 9, dove 1 e 9 indicano rispettivamente il la minima e la massima *rilevanza* [o *originalità/innovatività*, o *rilevanza/impatto internazionale*]» (*ibidem*). Inoltre per ciascun criterio viene proposto un campo libero facoltativo (con un'estensione massima di 1000 caratteri) in cui riportare «in formato libero la spiegazione dei punteggi» (*ibidem).*

La scheda di valutazione dell'Area 7 non presenta una serie di domande, ma una definizione di ciascun criterio estremamente simile a quella riportata nel bando (Anvur 2011, p. 7). In questa seconda formulazione non è presente alcun campo libero per la giustificazione dei punteggi assegnati o la registrazione dei commenti dei *referee;* le modalità di risposta presentano una etichetta semantica e sono riportati in modo da sottolineare il livello del punteggio (mentre nel caso dell'Area 13 avevano una parziale autonomia semantica):

«Alto 9 – Eccellente
8 – Notevole
7 – Distinto
Medio 6 – Molto buono
5 – Buono
4 – Soddisfacente
Basso 3 – Corretto
2 – Marginale
1 – Irrilevante» (*ibidem*)[72].

Le schede originariamente proposte dall'Anvur si presentavano dunque come strumenti strutturati e standardizzati, composti da tre *item*, uno per ciascun criterio, ciascuno richiedente l'assegnazione di punteggi (in una scala da 1 a 9 o in una scala da 0 a 3) e da almeno un campo aperto, facoltativo. Non è possibile conoscere quali e quante modifiche siano state apportate alla scheda originale dai singoli GEV o sub-GEV, ma almeno nel caso specifico dell'Area delle Scienze Politiche e Sociali il GEV ha utilizzato delle etichette semantiche piuttosto estese per ciascuna modalità di risposta (§ 4.1.2)[73].

La trasparenza sembrerebbe maggiore, su questo punto, in riferimento alla procedura di valutazione bibliometrica. In quasi tutte le Aree, infatti, per i prodotti ritenuti valutabili secondo questo approccio sono stati presi in considerazione due indicatori: uno riferibile al numero di citazioni ricevute, l'altro all'impatto delle riviste in cui il prodotto è collocato (l'*impact factor* di WoS,

[71] Traduzione dall'originale in lingua inglese, così anche per le citazioni a seguire.

[72] E' interessante notare che, trattandosi di un Area bibliometrica, nella definizione dell'internazionalizzazione viene richiesto esplicitamente di tenere conto dell'eventuale collocazione del prodotto in una rivista con un elevato *impact factor*.

[73] Dalle interviste emerge che l'Anvur offrisse ai GEV la possibilità di scegliere liberamente tra le due scale, ma non è chiaro quanta parte della formulazione del contenuto delle schede fosse a discrezione dei singoli GEV: «ciascun GEV ha scelto la propria scheda di valutazione, la differenza era semplicemente nel fondo scala… c'erano tre criteri, a ciascun criterio era associato un punteggio che andava, se ricordo bene, o da 0 a 3 o da 1 a 9» e le etichette semantiche «alcuni GEV le hanno poi date» (Intervista Benedetto). E' stato tuttavia sottolineato che «non rimanevano molti gradi di libertà» (Intervista Blasi).

l'indice *SCImago Journal Ranking*, noto come SJR, di Scopus, e simili)[74]. Non mancano però leggere differenze tra le Aree (Tabella 15). Il GEV 13, di Scienze economiche e statistiche, ad esempio, ha utilizzato per la classificazione delle riviste non solo l'*impact factor*, ma anche l'*impact factor a cinque anni*, l'*article influcence score* e una versione dell'*h index* calcolata per rivista e non per autore (Anvur, GEV 13, Appendice C). Inoltre, in quest'Area l'indicatore relativo al numero delle citazioni non viene considerato per tutti i prodotti, ma solo per gli articoli con un numero significativo di citazioni nelle riviste indicizzate in WoS nel periodo 2004-2010 (in rapporto agli anni trascorsi dalla pubblicazione). Il GEV 1 di Scienze matematiche e informatiche, invece, come già segnalato (*cfr.* § 1.3.3.2), ha utilizzato, oltre alle banche dati WoS e Scopus anche MathSciNet e al suo interno ciascun Sub-GEV ha scelto autonomamente le procedure di classificazione delle riviste anche utilizzando indicatori differenti: *impact factor*, *impact factor* a due anni, *impact factor* a cinque anni, o l'indice MQC fornito da MathSciNet.

Tabella 15 – Database e indici di riferimento per la valutazione bibliometrica per Area

Area		Database	Indici	
			Rivista	Articolo
Area 1	Matematica e scienze informatiche	Web of Science ; Scopus; MathSciNet	A seconda del sub-GEV: IF, IF a 5, SJC, MCQ (con differenti algoritmi classificatori)	Citazioni dalla data di pubblicazione al 31/12/2011
Area 2	Fisica	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 3	Chimica	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 4	Scienze della terra	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 5	Biologia	Web of Science	IF	Citazioni dalla data di pubblicazione al 31/12/2011
Area 6	Medicina	Web of Science	IF	Citazioni dalla data di pubblicazione al 31/12/2011
Area 7	Scienze agrarie e veterinaria	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 8	Ingegneria civile e architettura	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 9	Ingegneria industriale e dell'informazione	Web of Science ; Scopus	IF, 5YIF, AI; EF	Citazioni dalla data di pubblicazione al 31/12/2011
Area 11	Scienze storiche, filosofiche, psicologiche e pedagogiche	Web of Science ; Scopus	IF; SJC	Citazioni dalla data di pubblicazione al 31/12/2011
Area 13	Scienze economiche e statistiche	Web of Science ; Scopus; Google Scholar	5YIF; AIS; h-index	Citazioni 2004-2010 (utilizzato solo per far avanzare di una classe gli articoli con più di cinque citazioni, altrimenti ininfluente)

[74] Questa la spiegazione fornita dal coordinatore della VQR: «nella bibliometria ci sono stati proposti una quantità enorme di indicatori, dopo di che l'unica proxy affidabile sarebbe il numero di citazioni. Noi abbiamo scelto invece di usarne due, prima di tutto perché più ne usi meno gli indicatori sono suscettibili poi, come dire, di drogaggio successivo, sono meno un invito a farlo. In secondo luogo […] la valutazione indirizza comunque il lavoro di scrittura, dei giovani, tanto vale cercare di farlo in una maniera che consideriamo utile. Dire che si utilizza anche un indicatore, una proxy di qualità del contenitore noi riteniamo che sia una cosa utile, che sia soprattutto indispensabile nel momento in cui la valutazione si fa in anni molto vicini a quelli in cui si sono pubblicati i lavori, perché per un lavoro pubblicato da un anno, un anno e mezzo, in nessuna delle discipline considerate si può pensare che l'indicatore citazionale sia arrivato a regime» (Intervista Benedetto).

Se nel caso della valutazione bibliometrica è possibile dunque conoscere esattamente quali indicatori siano stati utilizzati da ciascun GEV o sub-GEV non è però disponibile alcuna argomentazione circa la loro selezione in relazione alle dimensioni concettuali a cui sono riferite. Si amplifica in questo passaggio il deficit di chiarezza nella definizione della qualità della ricerca evidenziato nel Capitolo 3. Se la definizione della qualità della ricerca e le sue dimensioni sono le stesse per la procedura di valutazione in peer review e per quella bibliometrica, se dunque i criteri alla base della valutazione sono i medesimi, i problemi relativi alla selezione degli indicatori risultano estremamente rilevanti dal punto di vista semantico, nonostante la relativa pubblicità, riproducibilità e controllabilità delle successive fasi del processo di operativizzazione.

E' importante qui sottolineare un punto decisivo: se da un lato la definizione operativa di qualità della ricerca utilizzata nella procedura di valutazione bibliometrica segue tutte le fasi della procedura standard maggiormente diffusa nelle scienze sociali, cioè il modello lazarsfeldiano, la procedura per la valutazione in peer review salta un passaggio fondamentale: quello della selezione degli indicatori. Di ciascuna dimensione concettuale viene infatti messa a punto e utilizzata una definizione operativa diretta, volta a rilevare lo stato del prodotto su ciascuna dimensione interrogando direttamente il revisore.

In altre parole se gli indici bibliometrici si configurano effettivamente come *indicatori* anche sul piano concettuale, le risposte fornite dai revisori sono *indicatori* sul piano operativo, dato che servono alla rilevazione di uno stato su una proprietà, ma non sul piano concettuale, dato che derivano da una definizione operativa diretta dei criteri in valutazioni vere e proprie.

Si tratta di una differenza essenziale e inevitabile: l'individuazione di concetti-indicatori di qualità della ricerca renderebbe infatti necessaria non solo la schematizzazione di tutte le possibili caratteristiche degli oggetti, ma anche la definizione di standard di qualità per ciascuna di esse. La complessità della classificazione che ne deriverebbe risulta difficile anche solo da immaginare: si pensi ai diversi tipi di prodotto, alle differenze nella struttura, nelle varie modalità di riferimento ai dati o alla letteratura, della centralità di ciascuna di queste proprietà e delle possibili combinazioni dei loro stati al fine della valutazione di un prodotto della ricerca. L'enorme varietà delle forme della comunicazione scientifica e delle loro possibili connotazioni qualitative rendono necessario l'intervento di un giudice: il parere di un soggetto esperto rappresenta una traduzione empirica enormemente meno complessa e rischiosa di qualsiasi classificazione basata esclusivamente sulle caratteristiche dei prodotti.

4.1.2 La scheda di valutazione dei prodotti nell'Area delle Scienze Politiche e Sociali

Il rapporto del GEV dell'Area di Scienze Politiche e Sociali, come già evidenziato, non riporta la scheda di valutazione, limitandosi più volte a indicarne alcune caratteristiche: il fatto che sia stata predisposta dal GEV stesso, che contenga tre domande a risposta multipla corrispondenti ai criteri indicati nel bando e un campo libero (Anvur 2013d, GEV 14, p. 19, p. 65, 73-74). Grazie alla disponibilità dell'Agenzia è stato possibile però ottenere un *fac simile* della scheda, in modo da poter

analizzare la sua struttura, la formulazione delle domande e delle modalità di risposta, i punteggi corrispondenti a ciascuna di esse e l'algoritmo di assegnazione delle classi di merito[75].

La scheda era costituita da tre domande in forma chiusa, a risposta multipla, che prevedevano quattro alternative di risposta ciascuna, più un campo libero, riservato ai commenti dei revisori, con una ampiezza massima di 500 caratteri (Riquadro 1).

Risultano evidenti tre differenze essenziali rispetto alle altre schede di valutazione che è stato possibile osservare, quella dell'Area 13 e quella dell'Area 7:

1. la formulazione della domanda è molto più sintetica;
2. ciascuna modalità di risposta presenta una articolazione semantica molto più estesa;
3. il numero delle modalità di risposta è ridotto da 9 a 4.

I primi due punti sono strettamente legati tra loro: se, da un lato, la formulazione della domanda non contiene nessuna specificazione del significato del criterio corrispondente, dall'altro, questa funzione sembra essere stata assegnata alla formulazione delle modalità di risposta, seppure con un esito non del tutto soddisfacente.

Il terzo punto, connesso agli altri due anche se meno intrinsecamente, ha delle conseguenze sul significato attribuibile a ciascuna modalità di risposta: dato che la scala ha così poche posizioni i prodotti possono essere classificati dai *referee* solo come pessimi, medio-bassi, medio-alti o ottimi rispetto a ciascun criterio. E' evidente la corrispondenza tra le modalità e le classi di merito, nondimeno la scelta di adottare quattro sole modalità di risposta influisce sulla *sensibilità* dello strumento, cioè sul rapporto tra il numero di modalità previste dalla definizione operativa e il numero di stati che la proprietà può assumere nella realtà (Marradi, 1980; 2007), dunque può influire sul grado di affidabilità del dato rilevato.

Vale la pena partire dalla questione centrale: quella della rispondenza della scheda alla definizione di qualità della ricerca. Ciascuna delle domande presenti nella scheda è riferita *direttamente* a uno dei tre criteri alla base della valutazione definiti nel bando: *rilevanza*, *originalità/innovazione* e *internazionalizzazione* (Anvur, 2011, p. 7; *cfr*. § 3.1).

[75] Nel corso delle interviste con il Presidente, i coordinatori sub-GEV e l'assistente dell'Area 14 non è stato possibile ottenere una descrizione precisa della scheda, i ricordi dei testimoni non apparivano chiari su questo punto. Nelle precedenti fasi del lavoro si è lavorato su una ricostruzione, basata su alcune interviste a revisori di Area 14 (si veda a questo proposito Fasanella e Di Benedetto, 2014). La disponibilità del direttore Torrini e delle dottoresse Blasi e Colizza (entrambe ex Assistenti GEV di Area 14), ha permesso di ottenere una copia della scheda (in data 27/04/2015) e di rivedere l'analisi effettuata a partire dalla ricostruzione, sulla base dei documenti ufficiali.

Riquadro 1 - Scheda di rilevazione di Area 14

D1. ***Rilevanza***:quale importanza ha il prodotto rispetto alla letteratura precedente?

a- ha influenzato e ampliato la conoscenza nel campo con contributi empirici e teorici importanti, mediante l'utilizzo di concetti, teorie, approcci, metodi e dati particolarmente impegnativi e convincenti.
b- analizza un tema/problema rilevante, mediante l'applicazione di concetti, teorie, approcci, metodi e dati consolidati.
c- analizza un tema/problema di portata limitata, mediante l'applicazione di concetti, teorie, approcci, metodi e dati convenzionali.
d- fornisce un contributo trascurabile alla conoscenza nel campo.

D2. ***Originalità/innovazione***: qual è il livello di originalità/innovazione del prodotto?

a- l'approccio, la metodologia e l'analisi sono molto originali e innovativi.
b- l'approccio, la metodologia e l'analisi sono solo parzialmente originali e innovativi.
c- l'approccio, la metodologia e l'analisi sono scarsamente originali e innovativi.
d- l'approccio, la metodologia e l'analisi non sono né originali né innovativi.

D3. /***nternazionalizzazione***: qual è il livello di internazionalizzazione del prodotto?

a- si posiziona (o si posizionerà) in modo molto significativo dal punto di vista dell'interesse e della visibilità internazionale.
b- si posiziona (o si posizionerà) in modo significativo dal punto di vista dell'interesse e della visibilità internazionale.
c- si posiziona (o si posizionerà) in modo scarsamente significativo dal punto di vista dell'interesse e della visibilità internazionale.
d- è (o sarà) irrilevante dal punto di vista dell'interesse e della visibilità internazionale.

Campo libero per commenti (massimo 70 parole o 500 caratteri).

E' possibile notare come la domanda relativa alla *rilevanza* non contempli, come invece la definizione dei criteri (Anvur, 2011), alcun riferimento alla ricaduta della ricerca in termini di congruità, efficacia, tempestività e durata, a meno che non sia interpretabile nel senso della congruità e della tempestività il riferimento alla portata del tema/problema analizzato. Solo nel primo e nel quarto *item* è individuabile chiaramente un riferimento diretto all'avanzamento della conoscenza nel campo, che però sostituisce quello alla portata del tema/problema d'indagine. Infine il riferimento a concetti, teorie, approcci, metodi e dati sembrerebbe legato alla dimensione del «valore aggiunto per l'avanzamento della conoscenza nel settore e per la scienza in generale» (*ibidem*). Rimandando per ora la discussione circa la formulazione dei singoli item, la domanda considerata nel suo complesso, da un punto di vista semantico propone al rispondente solo parte del significato che doveva essergli trasmesso. La scelta di attribuire un'importanza marginale alla

questione della ricaduta della ricerca può essere in parte giustificata dalla considerazione delle caratteristiche delle discipline afferenti all'Area, tuttavia la definizione operativa a causa di questa scelta presenta una copertura semantica parziale della dimensione concettuale a cui è riferita.

La domanda sull'*originalità/innovazione* fa riferimento esclusivamente all'approccio, alla metodologia e all'analisi, tuttavia stando alla definizione del criterio sarebbe da intendersi come «contributo all'avanzamento di conoscenze o a nuove acquisizioni nel settore di riferimento» (*ibidem*), non escludendo dunque contributi di natura tematica, concettuale e teorica. Nuovamente la copertura semantica offerta dalla definizione operativa risulta parziale, questa volta senza neppure la scusante relativa alle caratteristiche peculiari delle discipline dell'Area. Se è vero infatti che nelle Scienze Politiche e Sociali raramente la ricerca conduce a ricadute significative e durevoli è vero anche che una parte rilevante dei contributi scientifici in quest'Area sono di natura teorica e che anche gli studi empirici mirano essenzialmente a uno sviluppo in questo senso.

La domanda relativa all'*internazionalizzazione* sembrerebbe quella maggiormente aderente alla definizione del criterio, riferendosi tanto all'interesse, dunque alla rilevanza e all'apprezzamento della comunità scientifica, quanto alla visibilità, cioè alla competitività e alla collocazione editoriale. La questione relativa alla collaborazione scientifica risulta riferibile più alla pratica della ricerca che a un criterio di qualità dei suoi prodotti, e in effetti nel corso della VQR sono stati utilizzati indicatori di internazionalizzazione riferibili a questo aspetto[76] ma non per la valutazione della qualità dei prodotti della ricerca.

Il contenuto semantico della definizione operativa non è il solo fattore determinate per la forza di un rapporto di indicazione; almeno nel caso di una rilevazione tramite interrogazione non è infatti possibile valutare il rapporto di indicazione senza tenere conto dell'affidabilità della sua definizione operativa. La prima scelta importante nella progettazione della scheda di valutazione per i *referee* è stata quella di utilizzare domande in forma chiusa. La struttura di una domanda è in effetti in grado di influire sulle risposte (Pitrone, 1995; Pitrone, 2009), contribuendo a determinarle invece che limitandosi a registrarle (Campbell e Fiske, 1959).

Una forma strutturata dovrebbe assicurare una maggiore comparabilità delle risposte, evitando che il rispondente si riferisca ad aspetti, caratteristiche, eventi, ecc. irrilevanti dal punto di vista di chi ha formulato la domanda. Inoltre, in una forma chiusa «le alternative di risposta sono parte della domanda e in molti casi sono indispensabili alla comprensione della domanda stessa» (Pitrone, 2009, p. 154), aiutano ad ancorare i termini ambigui presenti nel testo della domanda e fissano il quadro di riferimento rilevante per chi ha costruito lo strumento, aiutando il rispondente a comprenderne il significato e dunque a fornire la risposta (Schwarz e Hippler, 1987; Schwarz et alii, 1991; Mauceri, 2003).

Il problema più evidente relativo alla formulazione delle modalità di risposta è la pluralità di oggetti e attributi a cui tutte fanno riferimento. Pitrone sottolinea come questo genere di domande

[76] In particolare per la collaborazione internazionale facciamo riferimento all'indicatore IRAS4.2, collegato al numero e alla qualità dei prodotti di ricerca con almeno un coautore straniero (definito come rapporto tra la somma delle valutazioni della struttura in un'Area ottenute dai prodotti eccellenti con almeno un coautore straniero e la somma delle valutazioni ottenute dai prodotti eccellenti con almeno un coautore straniero dell'Area, Anvur, 2013a), e l'indicatore R_{int}, che confronta la qualità media dei prodotti con almeno un coautore straniero della struttura con la media di area (il rapporto tra il punteggio medio ottenuto dai prodotti con coautore straniero della struttura in esame e il punteggio medio ottenuto da tutti i prodotti con coautore straniero nell'Area, Anvur, 2013a).

risultino «seriamente sotto-determinate», sia nel caso che i diversi oggetti a cui facciano riferimento siano in opposizione fra loro, sia nel caso contrario; in entrambi i casi infatti «l'intervistato si trova di fronte al dilemma se accettare o rifiutare in blocco i diversi oggetti proposti» (Pitrone, 2009, p. 217-218).

Negli item (b) e (c) della domanda relativa alla *rilevanza* si fa riferimento al tema (O1) o al problema (O2) analizzato e ai concetti (O3), alle teorie (O4), agli approcci (O5), ai metodi (O6) e ai dati (O7) applicati. Anche considerando insieme i primi due oggetti in quanto espressioni del contenuto e insieme gli ultimi cinque come espressione del metodo la formulazione della domanda continua a contenere più di un stimolo. Nell'item (a) manca il riferimento al tema/problema analizzato, sostituito da quello al contributo, teorico (O8) o empirico (O9), rispetto alla conoscenza nel campo. Infine nell'item (d) non è presente il riferimento a nessuno di questi oggetti specifici, solo al contributo nel campo (O10).

Gli oggetti inoltre non sono gli unici stimoli presenti nella domanda, vanno considerati anche gli attributi degli oggetti; ad esempio nella domanda relativa alla *rilevanza*:

- nell'item (a) si fa riferimento all'influenza (A1) e all'ampliamento (A2) del contributo, teorico (O8) o empirico (O9), rispetto alla conoscenza nel campo, alla importanza (A3) del contributo e alla grado di impegnatività (A4) e persuasività (A5) di concetti (O3), teorie (O4), approcci (O5), metodi (O6) e dati (O7);
- nell'item (b) si fa riferimento alla rilevanza (A6) del tema/problema (O1 e O2) e al grado di consolidamento (A7) di concetti (O3), teorie (O4), approcci (O5), metodi (O6) e dati (O7);
- nell'item (c) si fa riferimento alla portata (A8) del tema/problema (O1 e O2) e alla convenzionalità (A9) di concetti (O3), teorie (O4), approcci (O5), metodi (O6) e dati (O7);
- nell'item (d) il riferimento è alla trascurabilità (A10) del contributo nel campo (O10).

Il problema si presenta anche nelle altre due domande: nella seconda, relativa alla *originalità/innovatività* del contributo, si fa riferimento a tre oggetti (l'approccio (O1), la metodologia (O2) e l'analisi (O3)) e a due attributi (originalità (A1) e innovatività (A2)), e nell'ultima, relativa all'*internazionalizzazione*, gli oggetti sono due (interesse (O1) e visibilità (O2)) e due gli attributi (rilevanza (A1) e significatività (A2)). L'ambiguità relativa agli attributi è in effetti scarsamente rilevante nel caso dell'*internazionalizzazione*, infatti rilevanza e significatività sono sostanzialmente utilizzati come sinonimi, mentre nel caso dell'*originalità/innovatività* l'ambiguità è a monte della formulazione degli item, nella stessa etichetta assegnata alla dimensione concettuale. Un ultimo elemento di ambiguità, relativo esclusivamente alla domanda sull'*internazionalizzazione* è dovuto al riferimento a due diversi momenti temporali: il presente (è/si posiziona) e il futuro (sarà/si posizionerà).

La pluralità degli oggetti e degli attributi contenuti negli *item* rende discutibile l'assunto che tutti i revisori abbiano risposto allo stesso stimolo, in questo caso «è chiaro che la domanda diverrà non univocamente interpretabile dal momento che ci saranno soggetti che risponderanno facendo riferimento solo al primo oggetto menzionato, altri solo al secondo e, altri ancora, forniranno una risposta facendo riferimento ad entrambi gli oggetti (o asserzioni) considerati nel loro complesso» (Mauceri, 2003, p. 124).

La alternative di risposta proposte oltre che risultare univocamente interpretabili prese singolarmente, come insieme dovrebbero rispettare i tre requisiti logici della classificazione:

esaustività, *mutua esclusività* e *unicità del fondamentum divisionis* (Marradi, 1990a; Mauceri, 2003). Vale certamente la pena di approfondire ciascuna delle tre domande anche da questo punto di vista.

La domanda relativa alla *rilevanza* risulta nuovamente la più problematica:

- le modalità di risposta elencate non sono *esaustive*, ad esempio non è possibile segnalare che un contributo, pur trattando di temi e problemi rilevanti, abbia utilizzato metodi, concetti, dati ecc. convenzionali;
- gli item non sono *mutuamente esclusivi* nella misura in cui peccano in esaustività, in altre parole, nell'esempio già riportato il rispondente potrebbe trovarsi nella posizione di voler utilizzare l'item (b) i riferimento ai problemi trattati e l'item (a) in riferimento ai metodi;
- entrambi questi problemi dipendono dalla mancanza di un unico *fundamentum divisionis* alla base della costruzione degli item: da un lato infatti si vuole tener conto della rilevanza degli oggetti dell'indagine, dall'altro della adeguatezza dei suoi metodi (raggruppando davvero a grandi linee gli oggetti e gli attributi a cui si fa riferimento).

Una possibile soluzione a quest'ultimo problema sarebbe indubbiamente l'utilizzo di uno spazio degli attributi per la progettazione degli *item*, immaginiando dunque il caso in cui si voglia tener conto di due *fundamenta divisionis*: la rilevanza degli oggetti dell'indagine e l'adeguatezza dei suoi metodi. Finanche considerando entrambi gli attributi come dicotomici (Figura 3) salvo la contemplazione di riduzioni funzionali (che pure non parrebbero necessarie né opportune in questo caso) il rispetto dei due *fundamenta divisionis* richiederebbe la costruzione di quattro *item*.

Figura 3 – Spazio degli attributi: rilevanza degli oggetti dell'indagine e adeguatezza dei suoi metodi.

		Oggetti	
		Rilevanti	Non rilevanti
Metodi	Adeguati	(1)	(2)
	Non adeguati	(3)	(4)

Dalla formulazione degli *item* e dalla struttura generale delle domande sembra però che l'intento fosse quello di graduare i criteri in quattro modalità, in linea con le classi di merito, e in questo caso lo spazio di attributi darebbe luogo a 16 combinazioni (Figura 4).

Figura 4 – Spazio degli attributi: rilevanza degli oggetti dell'indagine e adeguatezza dei suoi metodi.

		Oggetti			
		Molto rilevanti	Rilevanti	Poco rilevanti	Non rilevanti
Metodi	Molto adeguati	(1)	(2)	(3)	(4)
	Adeguati	(5)	(6)	(7)	(8)
	Poco adeguati	(9)	(10)	(11)	(12)
	Non adeguati	(13)	(14)	(15)	(16)

Volendo dunque mantenere una sola domanda per ciascun criterio gli *item* nella domanda relativa alla *rilevanza* del contributo dovrebbero essere (almeno[77]) 16 per rispettare i tre requisiti logici della classificazione, non solo dunque l'unicità del *fundamentum divisionis*, costituito in questo

[77] Considerando separatamente i dieci oggetti a cui si fa riferimento nella domanda, pur continuando a immaginare di dover riferire l'attributo della rilevanza a cinque di essi e l'attributo dell'adeguatezza ai restanti cinque, e di graduare in quattro modalità ciascuno di questi attributi le combinazioni possibili sarebbero 4^{10}, cioè 1.048.576, decisamente troppe.

caso dallo spazio degli attributi, ma anche l'*esaustività* e la *mutua esclusività* delle modalità di risposta. Naturalmente una simile formulazione presenterebbe un tipo diverso di problemi, legati alla difficoltà per il rispondente di «valutare con uguale attenzione tutte le alternative» (Marradi, 1980, p. 56). Circa il numero massimo di alternative di risposta valutabili contemporaneamente non c'è accordo nella letteratura metodologica (Miller, 1956; Schutz, 1958; Galtung, 1967; *cfr.* Mauceri 2003), tuttavia dipende fortemente dal grado di complessità delle modalità stesse: «il rischio di distorsione più aumentare proporzionalmente anche in relazione alla difficoltà incontrato dal soggetto nell'interpretare ciascuna alternativa» (*ibidem*, p. 150).

Sarebbe dunque preferibile formulare una domanda distinta per ciascun oggetto, in questo caso dunque dovrebbero essere proposte (almeno) due domande: (almeno) una riferita alla rilevanza degli oggetti e (almeno) una riferita all'adeguatezza dei metodi. La formulazione di domande distinte renderebbe più semplice il riferimento di specifici attributi a specifici oggetti, semplificando notevolmente l'articolazione semantica delle alternative di risposta, ma complicherebbe le procedure di sintesi (*cfr*. § 6.1.1).

Le stesse problematiche si riscontrano anche in riferimento alla domanda circa l'*originalità/innovatività* del contributo:

- le modalità di risposta non sono *esaustive*, ad esempio non è possibile segnalare che un contributo pur con un approccio originale e/o innovativo abbia utilizzato una metodologia o un'analisi convenzionali;
- gli item non sono *mutuamente esclusivi* proprio a causa della loro mancanza di esaustività, in altre parole il rispondente potrebbe trovarsi nella posizione di voler utilizzare l'item (c) i riferimento all'approccio e l'item (a) in riferimento all'analisi;
- entrambi questi problemi dipendono dalla presenza di più *fundamenta divisionis*: anche considerando originalità e innovatività come indissolubili, approccio, metodologia e analisi dovrebbero rappresentare altrettanti *fundamenta divisionis*.

Di nuovo, voledo mantenere la formulazione di un solo quesito per ciascun criterio, il riferimento a uno spazio degli attributi appare la soluzione più semplice. Tuttavia uno spazio di attributi che tenga conto del grado di *originalità/innovatività* per ciascuno dei tre oggetti a cui si fa riferimento (approccio, metodologia e analisi) conduce a notevole moltiplicazione delle modalità di risposta. Graduando ancora il criterio in quattro modalità (etichettate come *nulla*, *scarsa*, *parziale* ed *elevata*, in conformità con l'attuale formulazione della domanda), lo spazio degli attributi presenta 16*4 (64) combinazioni (Figura 5), un numero troppo elevato perché una batteria di *item* risulti gestibile da parte del rispondente.

In questo caso dunque, come del resto nel caso in cui nella domanda sulla *rilevanza* si intenda mantenere il riferimento a 10 oggetti distinti, l'unica soluzione sarebbe proporre domande distinte per ciascun oggetto, anche in batteria, richiedendo per ciascuno di essi di indicare il grado di *originalità/innovatività.*

Figura 5 – Spazio degli attributi: originalità/innovatività di approccio, metodologia e analisi.

Originalità/innovatività della **metodologia**	Originalità/innovatività dell'**analisi**	*Originalità/innovatività* dell'**approccio**			
		Nulla	Scarsa	Parziale	Elevata
Scarsa	Nulla	(1)	(2)	(3)	(4)
	Scarsa	(5)	(6)	(7)	(8)
	Parziale	(9)	(10)	(11)	(12)
	Elevata	(13)	(14)	(15)	(16)
Scarsa	Nulla	(17)	(18)	(19)	(20)
	Scarsa	(21)	(22)	(23)	(24)
	Parziale	(25)	(26)	(27)	(28)
	Elevata	(29)	(30)	(31)	(32)
Parziale	Nulla	(33)	(34)	(35)	(36)
	Scarsa	(37)	(38)	(39)	(40)
	Parziale	(41)	(42)	(43)	(44)
	Elevata	(45)	(46)	(47)	(48)
Elevata	Nulla	(49)	(50)	(51)	(52)
	Scarsa	(53)	(54)	(55)	(56)
	Parziale	(57)	(58)	(59)	(60)
	Elevata	(61)	(62)	(63)	(64)

La domanda circa l'*internazionalizzazione* non è esente da queste stesse problematiche:

- neppure in questo caso le modalità di risposta risultano *esaustive*, ad esempio non è possibile segnalare che un contributo pur avendo una significativa visibilità a livello internazionale abbia ottenuto uno scarso interesse;
- questa mancanza di esaustività rende nuovamente gli item non *mutuamente esclusivi*, il rispondente potrebbe trovarsi nella posizione di voler utilizzare l'item (c) in riferimento alla visibilità e l'item (a) in riferimento all'interesse;
- nuovamente questi problemi dipendono dalla presenza di più *fundamenta divisionis*: anche considerando rilevanza e significatività come sinonimi i riferimenti alla visibilità e all'interesse dovrebbero costituire due distinti *fundamenta divisionis*; senza contare il doppio riferimento temporale presente nella formulazione delle modalità di risposta.

Sarebbe dunque opportuno riferire il grado della significatività/rilevanza distintamente alla visibilità e all'interesse internazionale. Di nuovo lo spazio di attributi produrrebbe 16 item (Figura 6).

Figura 6 - Spazio degli attributi: significatività/rilevanza della visibilità e dell'interesse internazionale.

		Visibilità internazionale			
		Nulla	Scarsa	Significativa	Molto significativa
Interesse internazionale	Nullo	(1)	(2)	(3)	(4)
	Scarso	(5)	(6)	(7)	(8)
	Significativo	(9)	(10)	(11)	(12)
	Molto significativo	(13)	(14)	(15)	(16)

Il doppio riferimento temporale presente nel testo della domanda complica ulteriormente le cose, adottato come ulteriore *fondamentum divisionis*, darebbe luogo a 64 combinazioni (Figura 7). In conseguenza di questo ulteriore elemento la formulazione di domande distinte per ciascun oggetto non appare come una soluzione praticabile, dato che ciascuna domanda andrebbe posta due

volte (una con riferimento al presente e una con riferimento al futuro) appesantendo la scheda e introducendo una ridondanza in grado di influire sulla qualità dei dati rilevati.

Figura 7 - Spazio degli attributi: significatività/rilevanza della visibilità e dell'interesse internazionale, nei due momenti temporali

Interesse internazionale		*Visibilità internazionale*							
		Presente				***Futura***			
		Nulla	Scarsa	Significativa	Molto significativa	Nulla	Scarsa	Significativa	Molto significativa
Presente	Nullo	(1)	(2)	(3)	(4)	(33)	(34)	(35)	(36)
	Scarso	(5)	(6)	(7)	(8)	(37)	(38)	(39)	(40)
	Significativo	(9)	(10)	(11)	(12)	(41)	(42)	(43)	(44)
	Molto significativo	(13)	(14)	(15)	(16)	(45)	(46)	(47)	(48)
Futuro	Nullo	(17)	(18)	(19)	(20)	(49)	(50)	(51)	(52)
	Scarso	(21)	(22)	(23)	(24)	(53)	(54)	(55)	(56)
	Significativo	(25)	(26)	(27)	(28)	(57)	(58)	(59)	(60)
	Molto significativo	(29)	(30)	(31)	(32)	(61)	(62)	(63)	(64)

Nessuna delle tre domande rispetta dunque i requisiti logici della classificazione, presentando un numero di modalità di risposta decisamente ridotto rispetto a quelle logicamente possibili. Non stupisce che la scheda «abbia creato alcune difficoltà di interpretazione, in molti casi segnalate dagli stessi *referee*» (Anvur 2013d, GEV 14, p. 65).

Si è detto che l'uso di quattro sole modalità, limitato all'Area 14 e alle Aree 1, 4, 5, 9 (tutte bibliometriche), riducendo il livello di sintesi e analiticità delle risposte può incidere sull'attendibilità dei dati. Come lo stesso Marradi sottolinea più volte (1992; 2007) è estremamente probabile che non si conosca il numero effettivo degli stati che la proprietà studiata può assumere, tuttavia «quale che sia il numero degli stati aggiungendo una categoria si aumenta la sensibilità, ed eliminandone una la si riduce» (2007, p. 107). E' vero che la vaghezza dell'informazione rilevata non è sempre da valutare negativamente e che un numero elevato di categorie di risposta può introdurre distorsioni di vario genere (Marradi, 1980), tuttavia in questo si è scelto di utilizzare il più ristretto «grado di approssimazione tollerato dal problema» (Campelli, 1996, p. 26)[78].

Nonostante, cioè, la coincidenza del numero di modalità con il numero delle classi di merito, l'espressione dei giudizi da parte dei revisori richiederebbe un livello di analiticità più elevato. Infatti il fine della rilevazione in questione è principalmente classificatorio e in vista di questo fine sarebbe stato forse preferibile adottare uno strumento più sensibile (come quelli delle Aree 13 e 7) dato che «è sempre possibile procedere a una riaggregazione degli stati di una variabile e, al contrario, sempre impossibile – a meno di non ripetere la rilevazione su quella specifica proprietà – abbassare il livello di sintesi con cui un'informazione è stata rilevata» (Mauceri, 2003, p. 73).

A questo proposito vale la pena rilevare che le domande proposte sono in effetti simili a *scale con categorie ordinate*, cioè scale cui «il *continuum* con cui si rappresenta la proprietà viene diviso in segmenti senza l'ausilio di un'unità di misura, ma semplicemente disponendo sul *continuum* stesso un certo numero di categorie» (Marradi, 1980, p. 59). L'assunto alla base delle domande di questo

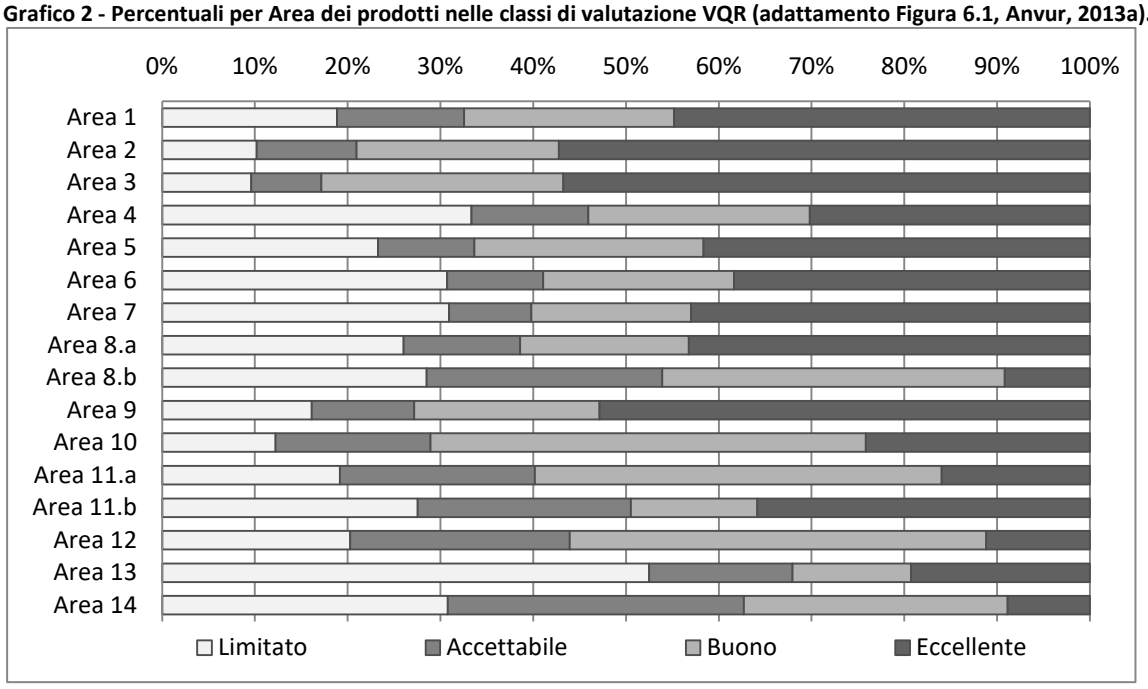
[78] Una osservazione di questo tipo è stata restituita anche dal presidente del GEV: «abbiamo scelto questo punteggio da 0 a 3 che forse… è stato sbagliato, se avessimo utilizzato una scala più lunga di punteggio questo avrebbe permesso di sfumare di più i giudizi» (Intervista Colozzi).

tipo è che il rispondente sceglierà la modalità di risposta più vicina alla propria posizione. Marradi (1980) ha criticato con forza questo assunto, sottolineando il fatto che «la distorsione è in linea di principio tanto maggiore quanto più è ridotto il numero delle categorie» (*ibidem*).

Le modalità di risposta proposte presentano inoltre una sostanziale autonomia semantica, una caratteristica che in genere risulta inversamente proporzionale all'ordinabilità (Marradi, 1980). In questo caso non sembra possibile sollevare dubbi sull'ordinabilità delle modalità di risposta, proprio in ragione del loro numero ridotto e delle formulazioni tendenzialmente estreme. In effetti, nel caso della domanda relativa all'*internazionalizzazione,* sembra poco plausibile che diversi rispondenti abbiano ordinato le modalità di risposta diversamente da: irrilevante, scarsamente significativo, significativo e molto significativo.

Purtroppo non è possibile esaminare dettagliatamente gli effetti delle definizioni operative proposte sull'attribuzione finale della classe di merito, tuttavia osservando la distribuzione delle classi di merito all'interno di ciascuna Area (Grafico 2) è possibile formulare alcune osservazioni.

Grafico 2 - Percentuali per Area dei prodotti nelle classi di valutazione VQR (adattamento Figura 6.1, Anvur, 2013a).



Innanzitutto è evidente che l'Area 14 presenta una quota molto ridotta di prodotti eccellenti, insieme all'Area 8b; inoltre, in confronto alle altre Aree ha una quota più elevata di prodotti valutati come buoni o accettabili. Infine la quota di prodotti limitati sembra abbastanza in linea con quella delle altre Aree prevalentemente non bibliometriche, pur risultando più alta della maggior parte dei corrispettivi nelle Aree bibliometriche.

E' un peccato non avere a disposizione neppure i dati distinti per approccio valutativo e bibliometrico, non potendo approfondire le semplici osservazioni appena formulate o tentare di fornirne delle spiegazioni. Ciò nonostante risulta estremamente plausibile che la formulazione della scheda abbia contribuito a limitare la quota di prodotti con una valutazione eccellente: le distorsioni possibili in relazione alle caratteristiche delle singole domande infatti non facilitano in nessun caso una valutazione positiva. La presenza di più oggetti nello stesso item, la mancanza di esaustività e

mutua esclusività delle modalità di risposta, ecc. incanalano i giudizi verso il basso: non ritenendo un contributo eccellente o buono per uno qualsiasi degli elementi riportati, anche se impeccabile o accettabile sotto gli altri punti di vista il rispondente è spinto a indicare la modalità che esprime una valutazione di livello inferiore. Ad esempio se si ritiene un prodotto originale, ma non «molto originale» sulla base delle modalità di risposta della domanda relativa alla *originalità/innovazione* si è spinti a definirlo come «solo parzialmente originale». Il numero molto piccolo delle modalità di risposta previste contribuisce a questo effetto. Se «riducendo il numero delle categorie si accresce la distanza media tra la posizione di un individuo e quella della categoria a lui più vicina» (Marradi, 1980, p. 61) e due delle categorie proposte presentano casi estremi (si pensi in particolare all'uso di espressioni come "trascurabile", "irrilevante", "importante", "molto") è del tutto plausibile immaginare un effetto distorsivo che spinga i rispondenti verso la categoria mediana[79].

Le possibili distorsioni generate dalla definizione operativa dei singoli criteri non sono però gli unici elementi critici nella rilevazione delle valutazioni della qualità dei prodotti nella VQR. Nel rapporto finale infatti si legge che al termine della compilazione della scheda «la classificazione era proposta al revisore per consentirgli di confrontarla con la definizione delle classi 1, 2, 3 e 4 della Sezione 2.5[80] e, eventualmente, di modificare i punteggi» (Anvur, 2013a, p. 26). Questa particolarità delle schede getta un'ulteriore ombra non solo sulla attendibilità, ma anche sulla validità della definizione operativa di qualità della ricerca: se infatti i *referee* possono modificare la classe di merito finale modificando i punteggi sui singoli criteri, perché non chiedere loro direttamente l'attribuzione della classe di merito? [81]

La qualità della ricerca è stata definita attraverso tre criteri, ciascuno tradotto in una domanda a risposta multipla: dunque sono stati selezionati e operativizzati gli aspetti del concetto ritenuti *rilevanti* ai fini dell'esercizio di valutazione. La specificazione degli aspetti rilevanti è un

[79] Il Presidente del GEV ha avanzato osservazioni simili, con spiegazioni in parte analoghe: «noi dopo economia siamo stati i più bassi, probabilmente è stato questo in parte dovuto, io credo, da una parte all'inesperienza, per cui le valutazioni… e a quel *misunderstanding* che riguarda l'internazionalizzazione […] Quindi abbiamo una sottostima dei lavori eccellenti dovuta a questo *misunderstanding* del concetto di internazionalizzazione. Forse se, oltre a questo, avessimo usato un punteggio un po' più ampio probabilmente avremmo alzato almeno la media, non i prodotti eccellenti ma almeno la media dei prodotti si sarebbe un po' alzata» (Intervista Colozzi).

[80] Si fa qui riferimento alla definizione delle classi riportata nel Bando del 2011: *Eccellente*: nel 20% superiore della «scala di valore condivisa a livello internazionale»; *Buono*: nel segmento 60% - 80%; *Accettabile*: nel segmento 50% - 60%; *Limitato*: nel 50% inferiore (Anvur, 2011, p. 7).

[81] Su questo punto i pareri dei membri dei GEV non evidenziano particolari problematiche: «in effetti su questo non abbiamo neanche dati, cioè non sappiamo quante persone abbiano poi corretto, dopo aver rivisto il punteggio, per cui la classe in cui andava a finire. Quindi questo è un dato su cui non abbiamo informazioni in realtà […] Devo dire, io ho la mia esperienza di *referee*, non è che mi sia capitato di avere una sorpresa nel momento in cui la scheda si chiudeva e mi dava la collocazione. Cioè il giudizio che avevo dato tutto sommato corrispondeva, e penso che sia stata una esperienza abbastanza condivisa, quindi non saprei dirle quanto è servito quel meccanismo di garanzia» (Intervista Colozzi); «nel momento in cui si deve chiudere una pratica e quindi una scheda, e quindi esprimere il giudizio sintetico conclusivo, è evidente che si va indietro e si va a verificare quello che si è scritto: che tipo di giudizio, che aggettivi, che avverbi, che verbi, che temi sono stati usati. Come in ogni operazione scientifica, come naturalmente tende ad essere la stessa valutazione: la rivedibilità è uno dei criteri fondanti. Nessuno può presumere di primo acchito di cogliere in pieno il senso di una frase, di un'intera opera, di un capitolo, di un articolo, quello che sia. Ha sempre, giustamente la possibilità di rifare il suo percorso di andare sui propri passi e di vedere se in effetti poi quello che alla fine si conclude è coerente rispetto al precedente» (Intervista Cipriani).

passaggio fondamentale: «un concetto è la sua intensione» (Sartori, 1984, p. 32), senza un riferimento agli aspetti rilevanti del concetto di qualità della ricerca si corre il rischio di rilevare un concetto di qualità della ricerca diverso per ciascun soggetto. Marradi ha evidenziato che non è possibile avere garanzie che:

«a) un concetto formulato da un pensante A sia esattamente uguale (quanto a intensione) a un qualsiasi concetto formulato da un pensante B, anche se A e B designano il loro concetto con lo stesso termine o espressione;

b) un concetto formulato dal pensante A al tempo *t* sia esattamente uguale a un altro concetto formulato dallo stesso pensante A al tempo *t'*, anche se A denomina i concetti allo stesso modo» (Marradi, 2007, p. 54).

Naturalmente non intendiamo portare queste osservazioni alle loro estreme conseguenze relativistiche, e a questo fine possiamo fare riferimento a quanto scritto da Campelli circa il principio di indicalità degli *accounts*: «la premessa di principio dell'assoluta indicalità prefigura in realtà la possibilità di spazi in cui l'indicalità stessa sia più o meno variabile, fino a risultare più o meno controllabile» (Campelli, 1991, p. 27).

Dal punto di vista della definizione operativa si intende dunque sostenere che tanto più il concetto è generale, astratto ed esteso, tanto più si corre il rischio di rilevare aspetti estranei della sua intensione utilizzando una definizione operativa diretta. Dare a un rispondente la possibilità di ritoccare le proprie risposte per modificare il loro esito classificatorio significa sostanzialmente dargli la possibilità di ri-considerare le dimensioni concettuali che ritiene rilevanti, anche se non lo sono affatto nella concettualizzazione di qualità della ricerca alla base del processo di valutazione. Essendo necessario ritoccare le risposte fornite alle tre domande per modificare l'assegnazione della classe di merito, inoltre, si inficiano deliberatamente la validità e l'affidabilità rispettivamente degli indicatori e dei dati relativi ai singoli criteri. La validità del rapporto tra la classificazione finale e il concetto è messa a rischio.

Lo stesso GEV 14 nella stesura del rapporto finale riporta alcune osservazioni significative: «è indubbio che questa scheda abbia creato alcune difficoltà di interpretazione, in molti casi segnalate dagli stessi *referee* al GEV 14. Tra i criteri di valutazione presenti il più problematico si è rivelato essere quello dell'internazionalizzazione, a causa della difficoltosa applicazione a un'area le cui ricerche in molti ambiti disciplinari rimangono legate all'ambiente culturale e al dibattito nazionale, e della sua operativizzazione che, probabilmente, non siamo riusciti a esplicitare in modo adeguato. Tale problematica si è tradotta perciò in una forte penalizzazione della maggior parte dei lavori pubblicati in italiano che, anche a causa della scala numerica adottata[82], ha probabilmente ridotto in modo significativo il numero dei prodotti che sarebbero stati considerati eccellenti sulla base degli altri due criteri e di un'interpretazione non restrittiva del criterio "internazionalizzazione"[83]. Su

[82] La questione della costruzione delle variabili, e dunque dei punteggi attribuiti alle modalità di ciascuna domanda presente nella scheda, è affrontata nel § 4.2.2. Il commento riportato è stato chiaramente confermato dal Presidente GEV (si veda la nota 78).

[83] Questo parere è confermato nelle interviste: «c'è stato un chiarissimo fraintendimento, ancora una volta soprattutto dagli italiani, sul parametro internazionalizzazione. Moltissimi non hanno capito il significato, nonostante fosse spiegato. Nella legenda della scheda c'erano quelle due righe che dicevano cosa va inteso per internazionalizzazione, invece, in automatico, moltissimi hanno detto: siccome è una rivista italiana non può essere internazionale quindi 0. L'80% dei prodotti in italiano ha avuto 0 perché erano in italiano, semplicemente perché erano in italiano, e questo li ha penalizzati molto. Primo problema: il fraintendimento

questo punto ci sentiamo di raccomandare al Ministero che nel prossimo Bando il criterio dell'internazionalizzazione sia meglio esplicitato facendo riferimento alle peculiari caratteristiche della ricerca svolta nelle diverse Aree» (Anvur 2013d, GEV 14, p. 65).

Se da un lato vengono sottolineati i limiti legati all'interpretazione dei criteri e della loro formulazione, dall'altro si riconosce l'utilità delle informazioni rilevate con il campo aperto presente nella scheda. Nel rapporto finale di Area si legge: «un altro suggerimento che ci sentiamo di proporre consiste nel rendere obbligatoria, nella scheda di valutazione, la compilazione del campo "commento" al giudizio. La presenza delle motivazioni, infatti, si è rivelata uno strumento molto utile per la convalida definitiva della valutazioni nella discussione interna ai *consensus groups*, mentre la sua assenza ha reso necessario in molti casi il ricorrere ad un terzo referaggio[84]» (*ibidem*).

L'opzionalità del campo aperto e la scarsità di indicazioni circa la sua compilazione gettano tuttavia un'ombra anche su questa caratteristica della scheda. Nell'Area 13 si chiedeva di riportare nel campo libero le motivazioni ai giudizi forniti su ciascuno dei tre criteri, nell'Area 14 non vi era alcuna indicazione circa la sua compilazione. Non solo la compilazione del campo ma anche il suo contenuto dunque erano completamente affidati alla discrezione del revisore.

---

del terzo criterio» (Intervista Colozzi); anche da membri del Consiglio Direttivo: «un altro tema su cui possiamo mettere l'attenzione è il tema dell'interpretazione da parte dei *referee* del concetto di internazionalizzazione, il quale va inteso correttamente come la capacità dell'autore e del suo prodotto di intervenire efficacemente all'interno della comunità scientifica internazionale, indipendentemente dalla lingua in cui si scrive» (Intervista Bonaccorsi); tuttavia l'interpretazione distorta del criterio non è del tutto estranea neppure ai membri del GEV: «a parte l'internazionalizzazione che costituisce un problema serio, perché implica la citazione, la traduzione eccetera… in realtà gli altri due criteri non dicono nulla, non sono niente» (Intervista Bazzicalupo).

[84] I pareri degli EV non sono uniformi circa l'effettiva affidabilità dei commenti riportati nel campo aperto: «sarebbe un rigo. Però sono perfettamente d'accordo. Ma non c'era neanche il tempo! Non si può utilizzare… è come nel referaggio dei lavori, degli articoli che ci arrivano continuamente per le riviste. Teoricamente c'è un giudizio, fatto articolato anche lì, più o meno uguale, sufficiente, eccetera... Poi c'è: commento. Un commento a un'opera ha un senso se entri in dialogo con l'autore, ma non è che uno può dire "sì, ma l'opera è…", non ha senso. A parte ora l'enorme quantità di prodotti, non era possibile perché non era chiaro cosa stavamo analizzando. Se un lavoro è scarso emerge dal voto basso, certo. […] nessuno l'ha fatto perché è talmente oneroso e faticoso scrivere una cosa del genere che non ha nessuna utilizzabilità, perché nessuno può leggerlo» (Intervista Bazzicalupo), ma dall'intervista al Presidente GEV emergono chiaramente le ragioni di queste osservazioni: «la scheda aveva una parte che poteva essere compilata, non obbligatoria: quello è l'errore: il non aver reso obbligatoria la motivazione del giudizio. Che cosa è successo… tenga conto che le percentuali che le dico non sono esatte, sono a memoria, però l'80% degli stranieri ha riempito la parte descrittiva della scheda dando la giustificazione del giudizio, il 90% degli italiani ha lasciato bianco. Quindi già questo ci ha permesso… siccome alla fine la VQR chiedeva che il giudizio finale non fosse quello di *referee*, ma fosse quello del GEV che doveva motivare, nel caso volesse dare un giudizio diverso da quello dei *referee*, eccetera. Allora per noi avere le ragioni del giudizio sarebbe stato estremamente utile a capire il perché si era valutato in un modo piuttosto che in un altro. Infatti nel caso degli stranieri, quando i lavori erano stati valutati da due stranieri, avendo la possibilità di mettere a confronto i giudizi che stavano dietro le due valutazioni, anche in caso di valutazioni diverse abbiamo potuto capire, per esempio, che c'era un giudizio identico a fronte di un punteggio diverso. In questo caso ha fatto fede il giudizio, non il punteggio, e abbiamo potuto mettere il punteggio che corrispondeva al giudizio. L'assenza di questo in quasi tutte le schede degli italiani ci ha reso molto più difficile il lavoro. Allora un suggerimento che noi abbiamo dato anche ai prossimi che dovranno fare la VQR è che questa parte della motivazione del giudizio deve essere obbligatoria nella scheda» (Intervista Colozzi).

.

E' evidente quanto questa scelta sia legata all'idea dell'assoluta uniformità e stabilità dei criteri e delle modalità di valutazione dei revisori, nonché alla condivisione di una specifica pratica. Un approccio più accorto avrebbe regolato maggiormente l'uso del campo libero, cercando di limitare la variabilità nel suo uso tra diversi revisori, ma anche dallo stesso revisore in riferimento a diversi prodotti. Date le sue caratteristiche alcuni revisori potrebbero aver compilato il campo per tutti i prodotti valutati, facendo puntualmente riferimento alle motivazioni per i punteggi assegnati su ciascun criterio, altri potrebbero aver utilizzato il campo solo per comunicare al GEV eventuali dubbi su alcuni prodotti, altri potrebbero aver inteso diversamente il suo scopo a seconda del prodotto in valutazione, altri ancora potrebbero non averlo utilizzato in nessun caso.

La mancanza di vincoli nelle condizioni d'uso non può che influire negativamente sulla completezza, la comparabilità e la validità delle informazioni rilevate tramite questo campo. E' chiaro che i commenti dei revisori avrebbero potuto essere utilizzati in tutti i casi nelle discussioni all'interno dei gruppi di consenso se fossero stati disponibili per tutte le revisioni. Inoltre, data la libertà di compilazione del campo, l'utilità del suo contenuto ai fini dell'assegnazione della classe di merito finale potrebbe variare sensibilmente a seconda dell'interpretazione che i revisori hanno dato al suo scopo.

4.1.3 Gli indicatori nell'Area delle Scienze Chimiche

Nell'Area delle Scienze Chimiche sono stati sottomessi 11.608 prodotti, il 97,8% dei quali è stato valutato tramite la procedura bibliometrica, la quota più elevata nel corso della VQR 2004-2010. Il numero delle valutazioni effettuate tramite analisi biblometriche e il fatto che le caratteristiche delle procedure utilizzate in quest'Area siano le più frequenti nella VQR (*cfr.* Tabella 15) ne fanno un caso studio ideale.

Il GEV di Scienze Chimiche ha scelto, come la maggior parte dei GEV delle aree cosiddette bibliometriche, di utilizzare due indicatori: uno legato al numero di citazioni ricevute da ciascun prodotto scientifico, l'altro all'impatto della rivista in cui il prodotto è stato pubblicato.

Prima ancora di affrontare la questione centrale del legame tra questi indicatori e le dimensioni concettuali della qualità della ricerca poste alla base della procedura di valutazione è opportuno fornirne una breve descrizione. L'intento è quello di esporre chiaramente non solo le procedure di calcolo, ma anche le principali argomentazioni riportate in letteratura pro e contro ciascuno degli indicatori utilizzati, prima di passare all'analisi del loro rapporto semantico con la qualità della ricerca e le sue dimensioni concettuali nella VQR.

Il numero di citazioni ricevute da un articolo viene in genere interpretato come un indicatore del suo impatto nella comunità scientifica, la sua definizione operativa è estremamente semplice (si tratta di un conteggio), ciononostante il suo utilizzo richiede una serie di scelte cruciali.

Perché si possa procedere al conteggio delle citazioni è innanzitutto necessario definire una finestra temporale in base ai propri obiettivi cognitivi. Ad esempio è possibile conteggiare le citazioni ricevute da un articolo a sei mesi dalla pubblicazione, per rilevarne l'impatto immediato, oppure quelle ricevute nel decimo anno dalla pubblicazione, per rilevarne l'impatto di lunga durata. Nel corso della VQR, per la maggior parte delle Aree[85], il periodo di riferimento utilizzato per il conteggio

[85] Fa eccezione l'Area 13, Scienze economiche e statistiche, che ha ristretto la finestra al 31/12/2010.

delle citazioni è quello incluso dalla data di pubblicazione dell'articolo al 31/12/2011 (cioè la data di aggiornamento dei dati acquisiti dall'Anvur per la conduzione della VQR; Anvur, 2011, p. 10 e Anvur, 2013a, p. 27; Anvur 2013d, GEV3, Appendice B, p. 20). In altre parole la finestra per il conteggio delle citazioni ha un ampiezza diversa a seconda della data di pubblicazione dell'articolo.

Naturalmente il conteggio delle citazioni richiede la disponibilità di una base di dati costruita *ad hoc*. I database citazionali più noti e vasti sono WoS di Thomson Reuters e Scopus di Elsevier, tuttavia ne esistono altri, più o meno centrati su singole discipline (si pensi a MathSciNet) e più o mano discussi (su tutti Google Scholar). L'Anvur, come già evidenziato, ha scelto di non legarsi a un solo database, utilizzandoli entrambi per il calcolo degli indici bibliometrici nella VQR. Ciò comporta naturalmente delle differenze non solo nel grado di copertura della produzione scientifica, ma anche nella struttura, nel tracciamento delle citazioni, nelle procedure di calcolo degli indici e così via (*cfr*. § 5.2.1).

In riferimento alla procedura utilizzata nel corso della VQR le informazioni disponibili sono poche, ma vi sono alcune questioni interessanti. Il conteggio delle citazioni nell'Area 3 (come in tutte le altre Aree bibliometriche), per ragioni di ordine tecnico, non ha escluso le autocitazioni (Anvur, 2013d, GEV3, Appendice, p. 20). L'algoritmo era tuttavia in grado di distinguere, calcolando distribuzioni cumulative empiriche separate per le citazioni, «gli articoli "scientifici" da quelli di rassegna, che ricevono notoriamente un maggior numero di citazioni» (*ibidem*, p. 24). Queste due questioni sono centrali nella letteratura scientometrica, e meriteranno dunque un approfondimento nel corso della discussione sul rapporto di indicazione.

Oltre all'indicatore citazionale riferito all'articolo la procedura prevedeva l'utilizzo di un indicatore riferito all'impatto della rivista ospitante. Gli indicatori selezionati per la VQR come indicatori di impatto sono diversi a seconda del database di riferimento: l'*impact factor* viene utilizzato per gli articoli indicizzati in WoS, mentre il SJR (*SCImago Journal Ranking*) viene utilizzato per gli articoli indicizzati in Scopus[86].

L'*impact factor* è «una misura della frequenza con cui l'"articolo medio" di un giornale è stato citato in un dato anno o periodo[87]» (Garfield, 1994). L'*impact factor* "classico" calcolato su JCR (*Journal Citation Report*) di WoS è il rapporto tra le citazioni ricevute nell'anno corrente e gli elementi citabili pubblicati nel corso dei due anni precedenti. Questo indicatore è in grado di eliminare alcune fonti di distorsione, come l'età, le dimensioni o la cadenza di pubblicazione della rivista, tuttavia non permette di controllare tutta una serie di altre questioni. Infatti il numero di citazioni ricevute non dipende esclusivamente dalla reputazione della rivista o dalla qualità dell'articolo, ma anche dalle dimensioni del campo di studio e dalla cultura citazionale della comunità scientifica di riferimento. La stessa Thomson Reuters mette in guardia contro un uso poco accorto dell'*impact factor*, sottolineando che non lo utilizza mai da solo per la valutazione di una rivista, e che dunque nessun altro dovrebbe farlo. In particolare i gestori di WoS sottolineano l'influenza del numero medio di citazioni effettuate da ciascun articolo pubblicato e la minore affidabilità dell'indicatore nei casi in cui il tempo trascorso dalla pubblicazione dell'articolo sia breve.

[86] Si noti bene che, in riferimento all'Area 3, l'indicazione dell'utilizzo dell'indice SJR come indice di impatto è riportata esclusivamente a latere di una figura nel rapporto finale di Area (Anvur, 2013d, GEV3, p. 91), e che più di una volta nel report finale e nei singoli report di Area si fa riferimento al JCR (*Journal Citation Report*, ad esempio Anvur, 2013a, p. 21), che è uno strumento offerto da WoS, non da Scopus, e che in effetti contiene l'*impact factor*.

[87] Traduzione dall'originale in lingua inglese.

Sono noti diversi problemi in riferimento all'affidabilità dell'*impact factor*. Ad esempio Thomson Reuters non fornisce una definizione accurata di "elementi citabili" (Moed e van Leeuwen, 1996). Calcolando il numeratore dell'IF WoS conteggia le citazioni a tutti i documenti, mentre al denominatore include esclusivamente gli articoli, le note e le review, escludendo editoriali, lettere, ecc. che pure risultano molto frequenti in alcune riviste (*ibidem;* Moed e Visser 2008; 1999). Dunque il numero medio di citazioni per articolo è in realtà distorto dalla considerazione di tutte le citazioni ricevute dalla rivista, ma solo di alcuni degli elementi effettivamente pubblicati. Questa particolare distorsione risulta tanto più grave dal momento che influisce in misura diversa in base alla struttura delle riviste e alla cultura citazionale della comunità scientifica di riferimento. In altri termini, ad esempio, se in un dato campo vi è un acceso dibattito e la comunità utilizza interventi su una specifica rivista per comunicare in proposito, citando molti contributi dello stesso genere, l'indice considererà al numeratore un gran numero di citazioni non riferite a elementi conteggiati al denominatore (Glänzel e Shoepflin, 1995; Moed e Visser, 2008; Vanclay, 2009).

Naturalmente vi sono diversi fattori in grado di influenzare il valore di questo indicatore. Il volume degli articoli pubblicati in una rivista, il formato degli articoli non citati in un certo periodo di tempo e la distribuzione di frequenza delle citazioni risultano fortemente correlate al valore dell'*impact factor* (van Leeuwen e Moed, 2005), ma solo per quest'ultima può essere individuata una relazione semantica con la qualità delle pubblicazioni. In altri termini l'indicatore in questione presenta uno scarso grado di affidabilità, dato che gli esiti della sua definizione operativa non registrano fedelmente gli stati effettivi degli oggetti sulla proprietà che intende rilevare (Marradi, 1980)[88].

L'indicatore considerato come corrispettivo dell'*impact factor* utilizzato nell'Area 3 per rendere conto dell'impatto della rivista in Scopus, il SJR, non è poi tanto «analogo» (Anvur, 2013a, p. 21) all'indice di WoS, poiché tiene conto non solo delle citazioni, ma anche del prestigio scientifico delle loro fonti. L'indice viene calcolato sulla base del numero delle citazioni ricevute in un dato anno, pesandole per il prestigio della loro fonte, rapportato al numero di pubblicazioni nei tre anni precedenti. Una citazione non vale uno in questo sistema, ma le viene assegnato un peso proporzionale al prestigio della rivista da cui proviene (l'algoritmo è simile a quello utilizzato da Google per l'ordinamento delle pagine). Il prestigio di una fonte in un dato anno viene equamente suddiviso tra tutte le citazioni che effettua in quell'anno. Questo algoritmo di calcolo permette di correggere alcune distorsioni che invece non vengono considerate nel calcolo dell'*impact factor*. In particolare tanto maggiore è il numero di citazioni effettuate da una fonte in un certo anno tanto minore sarà il loro peso, dunque corregge almeno in parte le distorsioni legate alle dimensioni e alla cultura citazionale di diversi campi di studio.

[88] Le preoccupazione dell'Anvur a proposito della affidabilità degli indicatori bibliometrici sembrano limitate alla loro manipolabilità, ad esempio con riferimento all'impact factor: «nell'impact factor delle riviste il dato è ben noto e l'errore, come ISI documenta, si riferisce a quei pochissimi casi l'anno di riviste, che vengono addirittura espulse da WoS, che manipolano il proprio impact factor e per adesso la letteratura suggerisce che siano pochissimi casi, anche se è una letteratura che come ovvio viene gonfiata un po' per sostenere che l'impact factor è inaffidabile. Nel senso che ISI ha una procedura interna per identificare le possibili anomalie, crescite improvvise e ingiustificate e i casi di manipolazione documentati sono effettivamente pochi. Quindi come dico spesso anche nelle varie conferenze il fatto di scoprire dei singoli casi di anomalie, come la rivista che gonfia artificialmente l'impact factor, è un esempio opposto, un esempio di quanto siano robuste le comunità scientifiche che riescono a detectare la manipolazione in maniera abbastanza rapida. Questo proprio dal punto di vista bibliometrico» (Intervista Bonaccorsi).

Nello schema sinottico che segue (adattato da Colledge *et al.* 2010, p. 219), sono riportate le principali caratteristiche dell'*impact factor* e del SJR. I due indicatori (IF e SJR) non sono progettati al fine di rendere conto dello stesso concetto e gli algoritmi di calcolo presentano delle differenze significative, tuttavia in letteratura sono rinvenibili contributi che ne attestano una buona convergenza (con una correlazione di 0.915; *cfr.* Russeau, 2009, p. 5). Ciò non toglie che questi due indici non possono essere considerati intercambiabili, né a livello procedurale né a livello concettuale, e che dunque sarebbe necessario valutare il loro rapporto di indicazione con lo stesso concetto di qualità della ricerca e controllare gli eventuali effetti distorsivi che potrebbero essere introdotti utilizzando uno di essi piuttosto che l'altro[89].

Tabella 16 – Presentazione sintetica delle principali caratteristiche degli indici IF e SJR (adattamento da Colledge er al, 2010, p. 219).

Caratteristica	***SJR***	***IF***
Database	Scopus	Web of Science
Finestra delle pubblicazioni	3 anni	2 (o 5) anni
Finestra delle citazioni	1 anno	1 anno
Inclusione delle auto citazioni della rivista	La percentuale delle autocitazioni è limitata a un massimo del 33%	Si
Normalizzazione per campo disciplinare	Si, il prestigio è distribuito su tutte le citazioni, tenendo conto delle differenti frequenze delle citazioni tra i campi	No
Delimitazione del campo disciplinare	Non richiesta dalla metodologia, intrinseca nel calcolo	No
Tipo di documenti usati al numeratore	Solo sottoposti a peer review: articoli, contributi a convegni, rassegne	Tutti
Tipo di documenti usati al denominatore	Solo sottoposti a peer review: articoli, contributi a convegni, rassegne	Esclusivamente "fonti": articoli, contributi a convegni, rassegne
Ruolo della fonte delle citazioni	Pesa le citazioni sulla base del prestigio della rivista che le effettua	Nessun ruolo
Effetto dell'inclusione di più rassegne	Dipendente dal prestigio delle riviste che citano le rassegne	Le rassegne tendono a essere più citate degli articoli sui risultati di ricerca, incrementando così il valore dell'indicatore
Effetti dell'allargamento della copertura dei database	Il database ha un prestigio fisso. Il prestigio viene diviso tra più riviste e così ridistribuito in modo tale che più prestigio sia rappresentato nei campi in cui il database è più completo	Complessivamente incrementa il valore dell'indice perché più citazioni sono presenti nel database. Non vi sono correzioni per le differenze nella copertura del database tra i campi disciplinari

[89] Nel corso delle interviste non è stato possibile approfondire le differenze tra lo SJR e l'impact factor e le loro possibili conseguenze, gli EV dichiaravano di non conoscerle approfonditamente, oppure sottolineavano l'ininfluenza del problema data la modalità di assegnazione delle classi: «sinceramente non lo ricordo perché siccome una volta presa questa decisione…comunque in realtà se non si usa l'impact factor assoluto della rivista ma il decile all'interno di ognuna categoria, quest'effetto è molto minore che se uno usasse in assoluto l'impact factor» (Intervista Barone).

In letteratura non mancano contributi in quest'ultima direzione. Ad esempio, Gonzalez-Pereira e i suoi colleghi (2010) hanno confrontato il SJR con un *impact factor* costruito *ad hoc* su una finestra di tre anni, concludendo che nonostante nel complesso la correlazione tra i due indici risulti elevata[90], il prestigio risulta concentrato in un numero inferiore di riviste e vi sono significative differenze tra i ranking ottenuti. Nella stessa direzione Elkins e il suo gruppo (2010) hanno sottolineato che nonostante l'elevata correlazione (0.89) accade che riviste classificate nel 10% più alto della distribuzione su indice risultano nel 10% più basso sull'altro. Le possibili distorsioni nel ranking includono: «differenze nella definizione degli indici, differenze nei record dei data set che contribuiscono al calcolo degli indici, differenze nei periodi di tempo su cui i ranking vengono calcolati, e gli errori[91]» (Elkins *et al.* 2010, p. 89).

Le questioni tecniche circa la costruzione degli indicatori, le caratteristiche dei database, le possibili distorsioni derivanti dalle procedure utilizzate non devono oscurare la questione centrale: la validità degli indicatori in rapporto agli obiettivi cognitivi di chi li utilizza. A questo proposito Moed (1996) sottolinea che «bisognerebbe tenere a mente che la bibliometria non è meramente una questione di procedure tecniche e gestione di dati. La validità degli indicatori bibliometrici è in gioco in quasi tutti i dibattiti o le dispute nel nostro campo, anche se in molti casi la discussione sembra focalizzarsi su questioni "tecniche". Evidentemente queste dispute sulle questioni connesse alla validità non possono essere semplicemente messe da parte e non può essere suggerito di decidere quale metodologia adottare senza mettere in questione la sua validità[92]» (1996, p. 189).

Stando alle definizioni operative e a quelle concettuali, apparentemente non vi è alcun legame *diretto* tra gli indicatori bibliometrici utilizzati e i tre criteri posti alla base della valutazione, piuttosto sembrerebbe che la loro scelta sia basata sul (tacito) presupposto che tanto più alto è il numero delle citazioni ricevute dall'articolo e/o il fattore d'impatto della rivista che lo ospita tanto più l'articolo è rilevante, originale/innovativo e internazionale. La definizione operativa della qualità della ricerca "scavalca" le sue dimensioni concettuali ed entrambi gli indicatori sono considerati in rapporto diretto con il concetto generale.

La rilevanza, definita come 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» (Anvur, 2011, p. 7), e l'originalità, definita come il «contributo all'avanzamento di conoscenze o a nuove acquisizioni nel settore di riferimento» (*ibidem*), sarebbero dunque indicate dal numero di citazioni ricevute e dalla collocazione del prodotto in una rivista con un elevato fattore di impatto. Il numero di citazioni e l'*impact factor* si configurano qui come veri e propri indicatori: misure *indirette* di proprietà non altrimenti rilevabili.

La questione è di cruciale importanza, dunque vale la pena di chiedersi se tra gli strumenti bibliometrici attualmente in uso vi siano indicatori con un legame semantico più stretto con la rilevanza o l'originalità dei contributi. La risposta è negativa[93]. La bibliometria permette infatti

[90] Il valore della r di Pearson riportato è simile a quello riportato in Russeau, 2009 (0.915, p. 5): 0.931 (Gonzalez-Pereira et al., 2010, p. 384).

[91] Traduzione dall'originale in lingua inglese.

[92] Traduzione dall'originale in lingua inglese.

[93] Le parole del Presidente Fantoni confermano questa impressione. Alla domanda circa la capacità delle procedure di cogliere i criteri, così come definiti dal ministero, la risposta è: «per quanto riguarda la bibliometria questo aspetto forse è un po' più complicato perché il livello citazionale risponde a questa domanda in due possibili modi, ora io parlo soprattutto per la fisica ché sono più esperto. Il livello di

l'analisi dell'impatto delle pubblicazioni (riviste o singoli articoli) sulla comunità scientifica, ma non fornisce strumenti in grado di valutare il contenuto dei prodotti, né dal punto di vista della rilevanza né da quello dell'originalità, per la semplice ragione che non è possibile individuare la motivazione alla base della citazione. Lo stesso Garfield evidenzia che indicatori come il numero di citazioni ricevute: «non dicono nulla circa la natura del lavoro, nulla circa le ragioni della sua utilità o del suo impatto. Questi fattori possono essere affrontati solo tramite l'analisi del contenuto del materiale citato e l'esercizio di un competente giudizio dei pari[94]» (1979a, p. 364).

Il legame con la dimensione concettuale dell'internazionalizzazione risulta decisamente meno lasco. La definizione di questa proprietà come il «posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica, inclusa la collaborazione esplicita con ricercatori e gruppi di ricerca di altre nazioni» (Anvur, 2011, p. 7), risulta infatti adeguatamente traducibile operativamente tramite i due indicatori selezionati. In particolare l'*impact factor* risulta un buon indicatore del «posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale», mentre il numero di citazioni risulta riferibile all'aspetto relativo all'«apprezzamento della comunità scientifica» (*ibidem*).

I due indicatori utilizzati per la valutazione dei prodotti risultano dunque semanticamente più vicini alla definizione ufficiale della dimensione dell'internazionalizzazione, ma vengono in effetti riferiti a tutti e tre i criteri. Si tratta di uno stiramento semantico: dal punto di vista concettuale questa opzione operativa si traduce con l'assunzione dell'internazionalizzazione come indicatore degli altri due criteri, mentre la struttura del concetto deducibile dalle definizioni proposte dall'Anvur tanto in riferimento ai criteri quanto in riferimento alle classi di merito, pone invece la rilevanza e l'originalità/innovazione in una posizione del tutto indipendente dall'internazionalizzazione (*cfr.* § 3.2). Non solo dunque la selezione degli indicatori prescinde completamente dalle dimensioni rilevanti del concetto, ma ne viola le relazioni reciproche.

Vale la pena di fare un passo indietro e di discutere l'assunzione del numero di citazioni ricevute come indicatore di qualità della ricerca a partire da una riflessione su cosa sia una citazione. Una definizione molto bella, ma semplice, è quella proposta da Cronin (1981), per cui le citazioni sono impronte congelate nel panorama dei risultati intellettuali, che danno testimonianza del passaggio delle idee.

---

innovatività e di originalità si potrebbe vedere dall'impact factor o da una cosa che ha a che fare con la natura della rivista, per dire: se io mando un lavoro a *Nature* o a *Physics Letters* che se non c'è originalità non te lo pubblicano […] D'altra parte magari può succedere che tu pubblichi un lavoro su *Nature* che ha un impact factor altissimo e il numero di citazioni che hai su quel lavoro è inferiore alla congruità dell'impact factor della rivista […] allora lì c'è questo problema di andare a vedere la congruità. Ora la citazione in linea di principio risponde a tutti questi principi, ma insomma la prassi quotidiana un po' meno perché certamente un lavoro molto innovativo può avere poche citazioni perché troppo innovativo, quindi hanno molte citazioni quei lavori di cui si sa già la risposta in buona sostanza, che vuoi una conferma. Questa è una patologia però della scienza, noi non possiamo farci molto su questo. Chiaro che avrei meno problemi se si trattasse di citazioni, cioè nei limiti estremi di tante citazioni o pochissimi citazioni direi che so che le cose funzionano, nel limite di citazioni che sono nella media, devo dire, penso che noi stiamo guardando più che altro se questo lavoro è nel *mainstream*, è congruo, è non sbagliato e dà alcune informazioni in più rispetto al passato; ma a questo grande livello di innovazione veramente non lo so» (Intervista Fantoni).

[94] Traduzione dall'originale in lingua inglese.

La questione è naturalmente molto più articolata di quanto possa apparire, e per comprenderne la complessità si assumerà brevemente il punto di vista della sociologia della scienza (per un rassegna più completa si veda Leydesdorff, 1998). Vi sono diverse teorie sull'utilizzo delle citazioni nella pratica della rendicontazione scientifica, tuttavia sono identificabili due filoni principali, uno associabile a una visione normativa del comportamento citazionale, strettamente legata alla concezione mertoniana della scienza, l'altro corrispondente a una visione più costruttivistica.

Secondo la teoria normativa gli scienziati citano i propri colleghi quando utilizzano il loro lavoro per progredire nella propria ricerca. La citazione qui si configura come un indicatore di influenza sul lavoro scientifico ma anche di riconoscimento del lavoro da parte di altri ricercatori. Nelle parole di Merton: «il riferimento svolge sia funzioni strumentali che simboliche nella trasmissione e nell'ampliamento della conoscenza. Strumentalmente, ci svela un lavoro che potevamo non conoscere prima, parte del quale ci può risultare di ulteriore interesse; simbolicamente, registra in archivi durevoli la proprietà intellettuale della fonte citata, fornendo un piccolo elemento di riconoscimento da parte dei colleghi della pretesa di conoscenza, accettata o espressamente rifiutata, che è stata fatta in quella fonte» (Merton, 1968, tr. it. 2000, p. 1198).

La visione costruttivista del comportamento citazionale è invece basata sulla sociologia della scienza di impostazione costruttivista (Latour e Woolgar, 1979; Knorr-Cetina, 1981) e sull'osservazione che il contenuto cognitivo degli articoli ha una scarsa influenza sulla loro ricezione. Dato che la conoscenza scientifica è socialmente costruita attraverso la manipolazione di risorse politiche e finanziarie, nonché attraverso espedienti retorici (Latour e Woolgar, 1979), le citazioni non possono essere soddisfacentemente descritte unidimensionalmente facendo riferimento al contenuto informativo dell'articolo, ma vanno contestualizzate politicamente, socialmente ed intellettualmente (Gilbert, 1977).

In sintesi «le citazioni si trovano all'intersezione tra due sistemi: un sistema retorico (concettuale, cognitivo) attraverso cui gli scienziati cercano di persuadersi a vicenda circa le loro rivendicazioni di conoscenza; un sistema delle ricompense (riconoscimento, reputazione), attraverso cui viene distribuito il credito per i risultati ottenuti[95]» (Cozzens, 1989, p. 440). Ciascuno di questi sistemi conduce a una diversa interpretazione delle citazioni: in relazione al sistema retorico queste sono interpretabili in termini di pertinenza, utilità e influenza, mentre la qualità e l'importanza del lavoro andrebbero collegate al sistema delle ricompense (*ibidem*).

Garfield nel 1962 classificò le citazioni i base alla loro posizione nel testo, al loro contenuto linguistico e alle variazioni, differenze e regolarità nei loro modelli di utilizzo, individuando una serie di "ragioni per citare" riconducibili sia alla teoria normativa che alla teoria costruttivista, in sintesi: rendere omaggio ai pionieri; attribuire credito per un lavoro correlato (omaggiare i pari); identificare metodologie, strumenti, ecc.; fornire una letteratura di sfondo; correggere il proprio lavoro; correggere il lavoro di altri; criticare lavori precedenti; consolidare rivendicazioni; segnalare prossimi lavori; fornire connessioni con lavori poco diffusi, poco indicizzati o non citati; convalidare dati e classi di fatti (costanti fisiche, ecc.); identificare le pubblicazioni originali in cui un'idea o un concetto erano trattati; identificare le pubblicazioni originali o altri lavori che descrivono un concetto eponimo o un termine; disconoscere il lavoro o le idee di altri (rivendicazione negativa); discutere le

[95] Traduzione dall'originale in lingua inglese.

rivendicazioni di priorità di altri (omaggio negativo). Il comportamento citazionale, come evidenziato da diversi studi (per una ricognizione della letteratura si vedano Bornmann e Daniel, 2008), è in effetti influenzato non solo dal contenuto degli articoli, ma anche da altri fattori, in parte non scientifici, il cui peso può variare fortemente da autore ad autore.

Le autocitazioni sono un perfetto esempio di quanto sia difficile distinguere le citazioni che effettivamente contengono una traccia del passaggio delle idee da quelle il cui scopo è slegato dal discorso scientifico. In altri termini: un autore può citare un proprio lavoro precedente perché questo costituisce la base del proprio lavoro attuale, perché vi è connesso dal punto di vista teorico o metodologico, o perché presenta delle evidenze utili all'argomentazione che si sta presentando, ma può farlo anche per auto-promuoversi. Non è raro che le autocitazioni vengano interpretate come una forma di doping per gli indici bibliometrici invece che come una normale pratica scientifica (a titolo di esempio: van Raan, 1998b; Brysbaert e Smyth, 2011). Il comportamento citazionale individuale può risultare anche fortemente deviante dai modelli proposti nella letteratura scientometrica (ad esempio Glänzel e Thijs, 2004; Thijs e Glänzel, 2006). Tuttavia «finora non sono state individuate tendenze allarmanti» e non sembrano esserci ragioni per escluderle dai conteggi (Glänzel *et al.* 2006, p. 275).

L'impossibilità tecnica di escludere le autocitazioni dai conteggi utilizzati per la VQR non dovrebbe dunque aver inciso significativamente sulla validità dell'indicatore citazionale come indicatore di impatto; resta tuttavia da capire se e quanto le citazioni possano effettivamente essere considerate indicatore valido in riferimento alla definizione di qualità della ricerca alla base dell'esercizio di valutazione.

Nella letteratura, purtroppo, non vi è accordo né sulle dimensioni della qualità della ricerca che il numero di citazioni sarebbe in grado di rappresentare, né sulla forza o la plausibilità di questo rapporto di indicazione. Glänzel e Schoeplin (1995) hanno proposto una interpretazione estremamente pragmatica, secondo cui, dal momento che le citazioni sono una delle principali forme di utilizzo dell'informazione scientifica, seppure non sono in grado di rendere conto della totalità del processo di ricezione, sono pur sempre un indicatore fondamentale. Van Raan (2005) sottolinea invece che le citazioni non sono in grado di fornire uno «specchio ideale» della performance scientifica, ma che possono essere utilizzate come indicatori di impatto in riferimento a gruppi di ricerca e su lunghi periodi. A queste condizioni, infatti, una interpretazione normativa del comportamento citazionale risulta meno debole che in relazione a singoli autori o prodotti, dato che più ampio è il gruppo e il lasso temporale di osservazione meno i comportamenti citazionali slegati dal discorso scientifico dovrebbero risultare in grado di influire significativamente sui risultati (ad esempio autoeludendosi).

In relazione agli articoli restano diversi fattori, slegati dal loro effettivo contenuto, in grado di influenzare il numero di citazioni ricevute per effetto delle caratteristiche dei processi comunicativi, ad esempio il numero di co-autori e il numero di citazioni effettuate (Glänzel e Thijs, 2004; Veiera e Gomes, 2009). L'effetto di questi fattori dal punto di vista del rapporto di indicazione non può che tradursi in un allargamento della parte estranea dell'indicatore.

Una di queste caratteristiche è il genere dell'articolo: è noto che gli articoli originali e le rassegne ricevono più citazioni rispetto a lettere, note e recensioni (Moed *et al.* 1999). Il fatto che nel corso della VQR in diverse Aree, tra cui quella di Scienze Chimiche, siano state calcolate distribuzioni distinte per le rassegne e gli articoli originali permette un certo grado di controllo sulle distorsioni

che potrebbero derivare da un loro confronto diretto, tuttavia alcune questioni restano aperte. L'accuratezza e l'affidabilità della classificazione dei documenti nei database bibliometrici non sono scontate (Harzing, 2013), inoltre la distinzione tra rassegne e articoli non è esaustiva: non è chiaro in quale distribuzione si collochino altri tipi di prodotti, come lettere, editoriali, recensioni. Inoltre i due database utilizzati classificano diversamente i documenti (si veda anche il § 5.2.1).

L'utilizzo di distribuzioni separate per review e articoli originali risulta particolarmente rilevante dal punto di vista del rapporto di indicazione. Il numero di citazioni è assunto allo stesso tempo come indicatore di rilevanza, originalità e internazionalizzazione. La separazione delle distribuzioni da un lato permette un maggiore controllo sulle dimensioni della rilevanza e dell'internazionalizzazione, limitando il confronto a elementi dello stesso genere, dall'altro è necessario riflettere sull'originalità. Il numero di citazioni può essere interpretabile come un indicatore di originalità in riferimento ad articoli o saggi, ma in riferimento alle rassegne il nesso con questa dimensione appare più labile. E' possibile infatti immaginare che una rassegna venga citata perché completa e accurata, dunque internazionale e rilevante, ma l'originalità è una dimensione difficilmente riferibile a questo genere di prodotto. Lo scopo delle rassegne infatti non è quello di proporre avanzamenti di conoscenza, ma quello di rendicontarli. Il rapporto di indicazione del numero di citazioni con una delle dimensioni concettuali, l'originalità, risulta dunque estremamente debole in riferimento alle rassegne, pur essendo plausibile per gli articoli, i saggi e i rapporti.

L'effettiva plausibilità del rapporto di indicazione tra gli indicatori di impatto delle riviste e la qualità dei prodotti della ricerca va considerata con attenzione. Oltre alle questioni semantiche già esposte, infatti, l'insidia più evidente in questo caso è la fallacia ecologica (Robinson, 1950; Leydesdorff, 2009). Lo stesso Garfield non manca di sottolineare che le riviste non sono entità omogenee (1971).

La questione può essere posta in questi termini: se un articolo viene pubblicato in una rivista con un impatto elevato (i cui articoli mediamente presentano un alto numero di citazioni, dunque sono molto considerati dalla comunità scientifica), non è detto che presenti le stesse caratteristiche degli altri articoli nello stesso contenitore (cioè che ottenga un elevato numero di citazioni). Lasciando perdere per un attimo gli indicatori, qui la questione è: se un articolo viene pubblicato in una buona rivista, sarà sicuramente un buon articolo?

In letteratura sono rinvenibili diversi contributi che contrastano l'utilizzo della valutazione della rivista al fine della valutazione dei prodotti che in essa vengono pubblicati (Seglen, 1997a; Seglen, 1997b; Starbuck, 2005; Jarwal, Brion e King, 2009). Nel caso specifico dell'*impact factor* Seglen (1997a) evidenzia l'asimmetria della distribuzione delle citazioni ricevute dagli articoli pubblicati in una stessa rivista, in grado di produrre sensibili distorsioni nel caso in cui si attribuisca ai singoli articoli il valore medio delle citazioni della rivista. Abramo *et al.* (2010) mostrano, con riferimento specifico all'Italia, che l'*impact factor* e le citazioni sono, nel complesso, fortemente correlate per intervalli di tempo estesi, ma poco correlati e in grado di determinare forti instabilità nei ranking per intervalli di tempo ristretti.

L'utilizzo di due diversi indicatori di impatto delle riviste implica naturalmente una riflessione supplementare: poiché sono diversi per costruzione e scopo è infatti necessario valutare il loro legame con il concetto che si intende misurare: la qualità della ricerca.

L'*impact factor* è progettato come un indicatore, appunto, di impatto delle riviste, mentre il SJR è progettato come un indicatore di prestigio. Leydesdorff (2009) ha analizzato diversi indicatori

bibliometrici riferiti alle riviste i un'ottica comparativa, cercando di comprendere se indicatori diversi risultino riferibili a diverse dimensioni concettuali. Tramite un'analisi in componenti principali Leydesdorff individua tre dimensioni in grado di sintetizzare il set di indicatori: l'impatto, le dimensioni e l'influenza. In sostanza riconduce sia l'*impact factor* che lo SJR alla dimensione dell'influenza, considerando quest'ultima come l'esito della combinazione tra le dimensioni e l'impatto della rivista. Una visione in linea con quella di Garfield, per il quale: «dovrebbe essere ovvio che l'influenza [di una rivista] è una combinazione dell'impatto e della produttività[96]» (1986, p. 445).

La conclusione è che: «nonostante la loro comprovata correlazione con gli aspetti legati alla qualità, le citazioni in generale e gli *impact factor* in particolare sono e restano principalmente indicatori di recezione della informazione scientifica. La possibilità di misurare la qualità scientifica di singole pubblicazioni attraverso le sole citazioni è un mito[97]» (Glänzel, 2008, p. 6).

4.2 Le variabili

Una variabile non è che un concetto operativizzato, cioè una proprietà rilevata attraverso una definizione operativa. Ciascuno dei criteri di valutazione della VQR una volta associato a una specifica definizione operativa può convertirsi in una variabile, e questo ulteriore passaggio contribuisce a determinare l'attendibilità del dato rilevato, ma anche la validità del rapporto di indicazione. Le scelte relative alla costruzione delle variabili infatti sono legate innanzitutto all'attendibilità, intesa come una «proprietà del rapporto fra il concetto che ha suggerito la definizione operativa e gli esiti effettivi che tale definizione prevede» (Marradi, 1980, p. 36). Nondimeno, nel valutare la validità del rapporto tra il concetto e l'indicatore è necessario tenere conto anche dell'attendibilità, «perché il termine ultimo del rapporto semantico con il concetto generale sono i dati che vengono registrati nella matrice grazie alla definizione operativa» (*ivi*, p. 37).

L'espressione *costruzione delle variabili* indica correntemente «l'insieme delle operazioni logiche e pratiche che conducono alla rilevazione e alla registrazione degli stati di un'insieme di oggetti cognitivi su una serie di proprietà» (Agnoli, 1992, p. 142).

A questo proposito Ammassari ha scritto che «il legame tra concetto e referente empirico ha […] il suo fulcro nella variabile e la sua scelta è determinante per la adeguatezza del processo di operazionalizzazione, come la validità della variabile è funzione di quello di concettualizzazione. Ma proprio per questo la validità non è data una volta per tutte» (1995, p. 183). Proprio in ragione dello stretto legame con il piano concettuale, da un lato, e con quello empirico, dall'altro, la costruzione delle variabili è una operazione complessa che implica uno sforzo «allo stesso tempo logico-concettuale e tecnico, rispetto al quale la rilevazione dei dati costituisce, spesso, solo una tappa intermedia (Agnoli, 1992, p. 143), questo soprattutto perché «la natura formale di una variabile non ha relazioni necessarie né con l'osservazione iniziale né con la corrispondente immagine concettuale (Lazarsfeld, 1958, p. 204).

Le variabili costruite nel corso della VQR, tanto nella procedura di peer review quanto in quella basata su indicatori bibliometrici, sono ordinali: metodologicamente, una sorta di *peggiore dei casi possibili*. Marradi ha trattato diffusamente delle distorsioni introdotte nelle definizioni operative

[96] Traduzione dall'originale in lingua inglese.
[97] Traduzione dall'originale in lingua inglese.

in riferimento alle variabili ordinali, dato che a suo parere in questo caso «i pericoli di distorsione sono talmente acuti da meritare una trattazione a parte» (1984, p. 59). L'uso di categorie ordinate costringe non solo, come già argomentato, a considerare insieme stati, più o meno diversi, sul continuum teorico proprietà, ma anche a decidere arbitrariamente la loro collocazione sul continuum al momento dell'attribuzione di un punteggio.

A rigore si ha un livello ordinale quando in assenza di una unità di misura l'ammontare della proprietà posseduto da un oggetto è confrontato con l'ammontare della stessa proprietà posseduto da un altro oggetto e l'esito del confronto è un giudizio di maggiore/uguale/minore (Coleman, 1964; Galtung, 1967). Nella pratica della ricerca è però pressoché impossibile confrontare tra loro tutte le coppie di casi (Borgatta e Bohrnstedt, 1980), dunque nelle scale definite come ordinali «sono le categorie, non i casi, ad essere confrontate in termini di maggiore o minore possesso della proprietà misurata» (Marradi, 1980, p. 54).

In genere alle categorie ordinate si assegnano dei codici numerici e questa operazione nasconde non poche insidie. E' fondamentale, innanzitutto, che i codici presentino una relazione monotonica diretta con l'ordine degli stati nella realtà, in secondo luogo bisognerebbe essere consapevoli del fatto che assegnare alle categorie codici nella forma di numeri naturali equivale considerarle tutte più o meno equidistanti tra loro (Marradi, 2007). Questo assunto tacito risulta particolarmente insidioso dato che è possibile che le categorie siano equidistanti, «ma nulla lo garantisce» (*ibidem*, p. 133). E' stato osservato più volte (Tufte, 1970, Marradi, 1980) che, se la nostra conoscenza sulla proprietà operativizzata ci suggerisce che le categorie non siano equidistanti, «assegnare i numeri naturali perché è oggettivo o meno arbitrario, introduce al contrario una distorsione; l'assegnazione dei codici "deve incorporare le conoscenze sostanziali che il ricercatore ha della proprietà in questione" (Tufte, 1970, pp. 440-441)» (Marradi, 2007, p. 133).

Troppo spesso nel ricorso alla serie dei numeri naturali «l'eguaglianza degli intervalli tra le categorie è solo una conseguenza meccanica, non programmata e non cercata» (Marradi, 1991b, p. 192). Tuttavia nel caso le etichette numeriche siano assegnate creando un rapporto monotonico con l'ordine delle categorie, la scelta dei codici numerici non ha effetti molto sensibili cui coefficienti che misurano la relazione tra le variabili (Labovitz, 1970; Marradi, 1980), purché il numero delle categorie sia sufficientemente ampio (O'Brien, 1985). Ancora una volta appare evidente che il rischio di distorsioni è tanto più elevato quanto più è ridotto il numero delle categorie.

4.2.1 Le variabili nella VQR

Un secondo punto d'ombra nella pubblicità, riproducibilità e controllabilità della procedura utilizzata nel corso della VQR per la valutazione dei prodotti è individuabile proprio nella costruzione delle variabili. A proposito della peer review*,* se i rapporti dell'Anvur sono poco trasparenti circa la scheda di valutazione, lo sono anche in riferimento alla attribuzione dei punteggi alle modalità di risposta. Nel rapporto finale dell'Agenzia è possibile individuare pochissime informazioni a riguardo: «a ogni risposta era associato un punteggio. La somma dei tre punteggi era confrontata con tre soglie per generare una classificazione finale in quattro classi» (Anvur, 2013a, p. 26). Né nel rapporto finale né nei singoli rapporti di Area vengono pubblicate le regole di sintesi e classificazione dei punteggi ottenuti dai prodotti (con alcune eccezioni *cfr.* § 4.1.2). Tanto i punteggi quanto le regole di

classificazione e sintesi sono stati però pubblicati dall'Anvur nel già citato breve documento on-line (Anvur, 2014). Le schede di valutazione hanno dunque prodotto punteggi veri e propri, considerabili come scale a intervalli.

Analogamente, riguardo alla costruzione delle variabili, nell'approccio bibliometrico diverse questioni di una certa rilevanza restano in secondo piano, tanto nel report finale quanto nei singoli rapporti di Area. Gli indicatori bibliometrici non sono stati utilizzati nella loro forma originale, ma riportati alle quattro classi di merito. La procedura di riconduzione alle classi prevedeva innanzitutto il calcolo delle distribuzioni cumulative dei due indicatori all'interno di una categoria disciplinare (le *Subject Category* di WoS e le categorie della *ASJC* di Scopus) per anno di pubblicazione, utilizzando le due basi dati complete, queste distribuzioni sono state poi suddivise in 4 classi contenenti una percentuale data di riviste (nel caso dell'indicatore di impatto) e di articoli (nel caso dell'indicatore citazionale) (Anvur, 2013a). La definizione quantitativa delle classi è quella riportata nel Bando del 2011, che fa riferimento alla segmentazione della «scala di valore condivisa a livello internazionale» (Anvur, 2011, p. 7):

e) *Eccellente*: nel 20% superiore della scala;
f) *Buono*: nel segmento 60% - 80%;
g) *Accettabile*: nel segmento 50% - 60%;
h) *Limitato*: nel 50% inferiore (*ibidem*).

Nel report finale della VQR non è esplicitato di quali soglie si tenesse conto per la riconduzione dei valori dei due indicatori bibliometrici alle classi di merito; questa informazione è presente nei report finali di Area o nelle relative appendici, di solito nell'appendice relativa ai criteri utilizzati dal GEV.

Quasi tutte le Aree hanno utilizzato le soglie teoriche, con l'eccezione dell'Area 9, Ingegneria Industriale e dell'informazione, in cui alcuni Sub-GEV hanno suddiviso in quartili la distribuzione delle riviste per impatto, e dell'Area 13, che ha utilizzato una procedura particolare sia per l'attribuzione delle classi di merito, sia con riferimento all'impatto delle riviste, sia relativamente alle citazioni ricevute dagli articoli (Anvur, 2013d, GEV 13, Appendice C).

Tabella 17 – Algoritmo di classificazione e principali caratteristiche della procedura per Area

Area	Nome	Algoritmo di classificazione	Distribuzioni separate per le rassegne	Categorie utilizzate	Articoli in riviste multisisciplinari
Area 1	Matematica e Scienze informatiche	le distribuzioni empiriche ordinate sono state divise in quattro classi (20%, 20% 10%, 50%, con alcune differenze per singoli sub-GEV)	No	Nuove categorie, proposte combinando SC e ASJC	Ricondotti ad altre categorie
Area 2	Fisica	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	No	SC; ASJC; PACS Numbers	Ricondotti ad altre categorie
Area 3	Chimica	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC; ASJC	Ricondotti ad altre categorie
Area 4	Scienze della terra	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC; ASJC	Ricondotti ad altre categorie
Area 5	Biologia	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC (con alcune aggregazioni)	Ricondotti ad altre categorie
Area 6	Medicina	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC (con alcune aggregazioni)	Ricondotti ad altre categorie
Area 7	Scienze agrarie e veterinaria	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	No	SC; ASJC	Valutati tramite peer review
Area 8	Ingegneria civile e architettura	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC; ASJC	Ricondotti ad altre categorie
Area 9	Ingegneria industriale e dell'informazione	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%) (ING/INF5 ha utilizzato la classificazione delle riviste di un sub-GEV di Area 1)	Si	SC; ASJC	Ricondotti ad altre categorie
Area 11	Scienze storiche, filosofiche, psicologiche e pedagogiche	le distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	Si	SC; ASJC	Ricondotti ad altre categorie
Area 13	Scienze economiche e statistiche	le citazioni erano ricondotte a due classi: >5 e <5; per le riviste distribuzioni cumulative sono state divise in quattro classi (20%, 20% 10%, 50%)	No	SC; ASJC	Analizzati come una categoria a sé

La decisione di non sfruttare l'intera informazione contenuta negli indicatori, ma di ricondurne i valori alle quattro classi di merito comporta naturalmente delle conseguenze. Lo svantaggio più evidente è relativo alla riduzione del livello di analiticità dell'informazione, e dunque della sensibilità della classificazione. Tuttavia in questo caso l'informazione "pura" relativa al valore dell'indicatore non sarebbe risultata conforme alle necessità informative dell'esercizio di valutazione, dunque ad una perdita di dettaglio dell'informazione corrisponde in questo caso un miglioramento nella sua rispondenza alle esigenze conoscitive in termini di comparabilità e sintesi.

4.2.2 Le variabili nell'Area delle Scienze Politiche e Sociali

Il rapporto del GEV dell'Area di Scienze Politiche e Sociali non solo non riporta la scheda di valutazione, ma neppure contiene indicazioni sulla scala utilizzata per la valutazione. L'unico punto del rapporto finale da cui è possibile intuire che vi sia stata una attribuzione di punteggi ai prodotti[98] è: «nel caso di valutazioni non convergenti dei revisori peer, il GEV ha creato al suo interno Gruppi di Consenso con il compito di proporre il punteggio finale del prodotto, oggetto del giudizio difforme dei revisori esterni, mediante la metodologia del *consensus report*» (Anvur 2013d, GEV 14, p. 19). Data però la disponibilità dell'Agenzia a fornire una copia della scheda e il documento circa le soglie per l'attribuzione delle classi di merito (Anvur, 2014), è stato possibile esaminare i punteggi corrispondenti a ciascuna delle modalità di risposta (Riquadro 2), oltre che l'algoritmo di assegnazione delle classi (*cfr*. § 4.3.2).

I punteggi attribuiti alle modalità di risposta sono identici per tutti e tre i criteri e vanno da un minimo di 0 a un massimo di 3. Anche nelle schede di valutazione dell'Area delle Scienze Agrarie e Veterinarie e dell'Area delle Scienze Economiche e Statistiche tutti e tre i criteri risultano operativizzati nello stesso modo: con l'attribuzione di un punteggio da un minimo di 1 a un massimo di 9 (*cfr.* Anvur, 2013d, GEV 13, Appendice E, pp. 112-113 e Anvur, 2013d, GEV 7, Appendice, p. 28).

[98] In riferimento alle strutture si parla di punteggio ottenuto, tuttavia in questo caso non ci si riferisce al punteggio assegnato ai singoli prodotti. Il punteggio ottenuto nella VQR è infatti dato dalla somma dei prodotti pesati: «sulla base del Bando VQR 2004-2010, ai singoli prodotti conferiti vengono assegnati pesi 1, 0.8, 0.5 e 0 a seconda che siano valutati Eccellenti, Buoni, Accettabili o Limitati; ai prodotti mancanti è assegnato peso -0.5, ai non valutabili è assegnato peso -1, e in casi accertati di plagio o frode si ha un peso -2» (Anvur, 2013d, GEV 14, p. 35).

Riquadro 2- Scheda di rilevazione Area 14, completa di punteggi

D1. ***Rilevanza***:quale importanza ha il prodotto rispetto alla letteratura precedente?

3. ha influenzato e ampliato la conoscenza nel campo con contributi empirici e teorici importanti, mediante l'utilizzo di concetti, teorie, approcci, metodi e dati particolarmente impegnativi e convincenti.
2. analizza un tema/problema rilevante, mediante l'applicazione di concetti, teorie, approcci, metodi e dati consolidati.
1. analizza un tema/problema di portata limitata, mediante l'applicazione di concetti, teorie, approcci, metodi e dati convenzionali.
0. fornisce un contributo trascurabile alla conoscenza nel campo.

D2. ***Originalità/innovazione***: qual è il livello di originalità/innovazione del prodotto?

3. l'approccio, la metodologia e l'analisi sono molto originali e innovativi.
2. l'approccio, la metodologia e l'analisi sono solo parzialmente originali e innovativi.
1. l'approccio, la metodologia e l'analisi sono scarsamente originali e innovativi.
0. l'approccio, la metodologia e l'analisi non sono né originali né innovativi.

D3. /***nternazionalizzazione***: qual è il livello di internazionalizzazione del prodotto?

3. si posiziona (o si posizionerà) in modo molto significativo dal punto di vista dell'interesse e della visibilità internazionale.
2. si posiziona (o si posizionerà) in modo significativo dal punto di vista dell'interesse e della visibilità internazionale.
1. si posiziona (o si posizionerà) in modo scarsamente significativo dal punto di vista dell'interesse e della visibilità internazionale.
0. è (o sarà) irrilevante dal punto di vista dell'interesse e della visibilità internazionale.

Campo libero per commenti (massimo 70 parole o 500 caratteri).

Nella traduzione di un indicatore in una variabile non bisognerebbe perdere di vista il rapporto di indicazione, cioè la sua relazione semantica con il concetto. Il punteggio corrispondente a ciascuna modalità è tanto più alto quanto maggiore è la corrispondenza del prodotto al criterio indicata dall'etichetta semantica. Appare inoltre corretto assegnare un punteggio 0 a modalità che corrispondono anche semanticamente all'assenza della proprietà, come nel caso di *originalità* e *internazionalizzazione*, anche se il legame tra etichetta e punteggio risulta meno preciso nel caso della *rilevanza*, dove il riferimento è a un "contributo trascurabile alla conoscenza nel campo".

I punteggi risultano effettivamente coerenti con il significato delle modalità di risposta associate, nondimeno, il fatto che nella scheda di valutazione le modalità di risposta erano etichettate con i rispettivi punteggi potrebbe avere influito sulla risposta fornita da parte del revisore. L'assegnazione di un punteggio 0 potrebbe essere stata percepita ad esempio come eccessivamente severa, spingendo i revisori verso i punteggi mediani, come anticipato nel § 4.1.2.

La soluzione adottata attribuisce a tutti e tre i criteri lo stesso peso nella determinazione della classe di merito (*cfr.* § 4.3.2). I punteggi presentano lo stesso campo di variazione e le tre

variabili hanno lo stesso ordine di scala, ma vale la pena sottolineare che una soluzione differente potrebbe risultare ugualmente legittima, o perfino auspicabile, dato che «considerare ugualmente validi tutti gli indicatori è un assunto indimostrabile esattamente come supporre il contrario» (Marradi, 2007, p. 189). I punteggi sono numeri naturali, dunque l'assunto alla base della costruzione delle variabili è quello dell'equidistanza tra le categorie ordinate. Un'equidistanza appunto assunta, ma non comprovata empiricamente né teoricamente fondata.

Esistono procedure pensate per la costruzione di scale per le quali non sia possibile assumere l'equidistanza tra le categorie. A titolo di esempio si pensi a una procedura tipo Thurstone, mirata a ordinare le modalità di risposta e a misurare, avvalendosi a questo fine del contributo di un gruppo di giudici, la distanza semantica tra una categoria e l'altra. In una procedura thurstoniana, ispirata alla tecnica degli *equal appearing intervals* (Thurstone, 1928), si potrebbe immaginare di attribuire a ciascuna modalità di risposta un punteggio rispondente alla sua posizione sul continuum della proprietà. In questo caso, se la distanza percepita tra la prima e la seconda modalità di risposta risultasse differente dalla distanza percepita tra la seconda e la terza e da quella percepita tra la terza e la quarta, sarebbe stato possibile optare per punteggi diversi da quelli utilizzati.

4.2.3 Le variabili nell'Area delle Scienze Chimiche

Nell'Area delle Scienze Chimiche era prevista l'assegnazione di una classe di merito a ciascun prodotto per ciascuno dei due indicatori utilizzati, sulla base della sua posizione nella «scala di valore condivisa a livello internazionale» (Anvur, 2011a, p. 7).

In sostanza tanto per il numero di citazioni ricevute quanto per l'*impact factor* viene calcolata la distribuzione cumulativa all'interno di ciascuna *subject category (SC)* o categoria dell'*all science journal classification (ASJC)*, per anno di pubblicazione, e al prodotto viene assegnata la classe corrispondente alla sua posizione in quella distribuzione, cioè se si trova nel 20% superiore della distribuzione ottiene la classe *eccellente*, *buono* se si trova nel 20% successivo della distribuzione, *accettabile* nel successivo 10% e infine la classe *limitato* viene assegnata al 50% più basso. La distribuzione cumulativa viene calcolata sugli interi database, considerando dunque l'intera produzione scientifica indicizzata in ciascuno di essi, senza limitare l'analisi alla produzione italiana.

La riconduzione in classi è in grado di "correggere" alcuni dei principali difetti dei due indicatori utilizzati, in particolare la variabilità delle distribuzioni tra campi di studio e nel tempo.

Già nel 1979 Garfield evidenziava come comparare il numero di citazioni in diversi campi di studio sia un'operazione impropria, dato che il «potenziale citazionale» può variare sensibilmente da un campo all'altro. Inoltre sottolineava l'esistenza di variazioni sensibili in relazione ad alcune caratteristiche delle citazioni: «quanto velocemente un paper viene citato, quanto tempo impiegherà il tasso di citazioni a raggiungere il suo picco e quanto a lungo il paper continuerà a essere citato[99]» (Garfield, 1979b, p. 248). In letteratura sono rinvenibili numerosi studi sia in riferimento alla variabilità delle caratteristiche delle citazioni in differenti campi di studio (van Raan, 2008a; Althouse *et al.* 2009; Costas *et al.* 2009), sia alla loro variabilità nel tempo (Wilson, 2007; Althouse *et al.* 2009; Olgden e Bartley, 2008; Neff e Olden, 2010).

[99] Traduzione dall'originale in lingua inglese.

In particolare la procedura risulta in grado di porre in parte rimedio alla variabilità del valore dell'*impact factor* o del SJR per anno e campo di studi. La procedura prevede infatti la costruzione di una distribuzione cumulativa distinta per categoria e anno, facendo sì che la classificazione di ciascun articolo non risenta delle differenze tra le culture citazionali tra i campi di studio o del volume complessivo delle citazioni in un determinato campo di studi entro un dato lasso di tempo.

Le differenze tra le specifiche culture citazionali dei diversi campi di studio vengono controllate anche in riferimento al numero di citazioni, grazie al calcolo di distribuzioni distinte per *subject category (SC)* o *ASJC*, ma la procedura non risulta altrettanto efficace dal punto di vista della normalizzazione per anno. Mentre infatti la finestra temporale di riferimento per il calcolo dell'*impact factor* e del SJR è fissa (nel primo caso due anni, nel secondo tre), le citazioni vengono conteggiate dalla data di pubblicazione dell'articolo al 31/12/2011, dunque la durata della finestra temporale è diversa per anno di pubblicazione.

In altri termini per gli articoli pubblicati nel 2004 la finestra temporale per le citazioni è di sette anni, per quelli pubblicati nel 2005 di sei, per quelli pubblicati nel 2006 di cinque, e così via fino a giungere a una finestra temporale di un solo anno per le pubblicazioni del 2010. Nonostante, dunque, la procedura sia identica a quella utilizzata per la riconduzione in classi dell'*impact factor,* nel caso delle citazioni non si ottiene una piena comparabilità della classificazioni, dato che queste sono ottenute sulla base di osservazioni di durata differente.

Sarebbe stato più corretto utilizzare una finestra temporale identica per tutti gli anni di pubblicazione, indipendentemente dall'ampiezza della finestra temporale su cui i dati risultano disponibili. Ad esempio per gli articoli pubblicati tra il 2004 e il 2008 sarebbe stato possibile conteggiare le citazioni ricevute nei tre anni successivi alla data di pubblicazione, mentre per gli articoli pubblicati nel 2009 e nel 2010 sarebbe stato possibile considerare solo le citazioni ricevute entro un anno dalla pubblicazione. Si tratta di una soluzione semplice e solo parziale, tuttavia questo accorgimento avrebbe almeno reso comparabili tra loro le classificazioni destinate a contribuire all'assegnazione della classe di merito finale tramite ciascuna matrice (*cfr*. § 4.3.3).

I presupposti su cui si basa la procedura di riconduzione degli indici alle classi sono semplici: «la "normalizzazione" degli indicatori in questo caso va intesa in modo inusuale: l'ordinamento o le quantità da esso derivate servono come base per la comparazione[100]» (Shubert e Braun, 1996, p. 318). Nelle parole di Garfield: «invece di comparare direttamente il numero di citazioni di, ad esempio, un matematico con quelle di un biochimico, entrambi dovrebbero essere classificati in un ranking con i loro pari, e la comparazione dovrebbe avvenire tra i ranking» (1979a, p. 367). In sostanza la classificazione degli articoli in base alla loro posizione nella distribuzione cumulata dell'indicatore si configura come una normalizzazione per campo di variazione e numerosità dei casi, sfruttando la loro posizione relativa (Shubert e Braun, 1996¸ Bornmann, 2013).

La rispondenza con la definizione delle classi di merito in termini di collocazione nella scala di valore condivisa a livello internazionale (Anvur, 2011) risulta evidente. L'estrema sintesi dell'esposizione dei livelli di merito in questo caso evita l'introduzione di elementi di ambiguità o vaghezza, permettendo l'individuazione di una definizione operativa adeguata alle definizioni proposte, sia dal punto di vista semantico che dal punto di vista procedurale.

[100] Traduzione dall'originale in lingua inglese.

Shubert e Braun (1996) hanno posto in evidenza il fatto che in genere nella valutazione è semplice distinguere l'eccellenza dalle performance al di sotto dello standard, la difficoltà sta nel tentativo di differenziare ciò che sta nel mezzo. E' qui che «i risultati sono generalmente fortemente dipendenti dalla procedura[101]» (Shubert e Braun, 1996, p. 317). In sostanza i ranking semplici (cioè la posizione puntuale dell'articolo nella distribuzione) funzionano benissimo per gli estremi delle distribuzioni, cioè per l'individuazione dei migliori e dei peggiori, mentre i ranking normalizzati (che riconducono cioè a un certo numero di classi le posizioni nell'ordinamento) funzionano meglio per i casi collocati al centro della distribuzione. Di qui l'opportunità di utilizzare classi di ranking non lineari (*ibidem*), cioè di non utilizzare soglie lineari, equidistanti (ad esempio i decili), ma di scegliere le soglie in base alla forma della distribuzione (Bornmann, 2013).

Le soglie percentuali utilizzate nel corso della VQR non sono equidistanti, tuttavia sono fisse, uguali per tutte le distribuzioni e determinate indipendentemente dalle loro caratteristiche. La mancanza di argomentazioni circa la scelta delle soglie non permette di sostenerne o contrastarne i presupposti, ma in questo caso l'ombra sulla trasparenza del processo di valutazione viene proiettata direttamente dal Ministero, il quale aveva fissato le soglie per le classi di merito già nel Decreto Ministeriale che indiceva la procedura (DM 17 del 15 luglio 2011)[102].

E' importante sottolineare che le distribuzioni degli indici di impatto e del numero di citazioni ricevute non seguono una funzione data. Nonostante la maggior parte delle più note "leggi" scientometriche faccia riferimento a distribuzioni esponenziali[103], nella realtà la forma delle distribuzioni può variare sensibilmente tra campi di studio e/o nel tempo. Non si tratta esclusivamente di una questione di scala (Radicchi *et al.* 2008), nonostante in molti campi la distribuzione risulti estremamente simile nella forma vi sono infatti delle eccezioni (Waltman *et al.* 2012). E' per questa ragione che Shubert e Braun sottolineano che seppure è possibile utilizzare le leggi bibliometriche per la creazione di schemi classificatori, è importante tenere a mente che queste sono «piuttosto approssimative, in alcuni casi addirittura contraddittorie[104]» (1996, p. 319)[105].

Inoltre esiste un problema legato ai pareggi: chiaramente se su dieci pubblicazioni una ha 50 citazioni (top 10%), due ne hanno 20 (20%), quattro ne hanno 10 (40%) e tre ne hanno 5 (30%) non sarà possibile suddividere la distribuzione cumulativa in classi da 20%-20%-10%-50%, ma sarà necessario "arrotondare", modificando di conseguenza l'ampiezza delle classi (Waltman *et al.* 2012; Waltman e Schreiber, 2012; Bornmann, 2013). Nei rapporti finali dell'Anvur non è possibile

[101] Traduzione dall'originale in lingua inglese.

[102] Il professor Bonaccorsi esprime insoddisfazione in relazione a questo aspetto: «ha toccato un tasto delicato [...] l'articolazione della scala, in particolare con la terza fascia del 10% che ha creato un problema metodologico serio perché nell'arco di 11 punti si passa da 0 a 0.8… dall'inferno al paradiso al purgatorio… con una variazione molto bassa. E' un problema di metrica sulle scale dove la robustezza ti impone, anche se non c'è la cardinalità ma un giudizio ordinale che diventa cardinale… richiede una qualche attenzione al modo con cui una variabile continua venga poi mappata su una discreta. Quello credo sia un errore francamente la cui origine a me sfugge, probabilmente anche ai colleghi. Forse è un'eredità della VTR precedente, o forse come spesso accade casualità nei decreti ministeriali» (Intervista Bonaccorsi).

[103] Si pensi alla legge di Lotka (1926), a quella di Bredford (1934), o a quella di Zipf (1949).

[104] Traduzione dall'originale in lingua inglese.

[105] Recentemente nella letteratura scientometrica la questione della forma delle distribuzioni in relazione alla normalizzazione tramite ranking è tornata alla ribalta (Egghe 2005, 2009; Mansilla et al., 2007), con un acceso dibattito non esente da contraddizioni e critiche (per tutti Waltman e van Eck, 2009; Egghe e Waltman, 2011).

individuare nessun riferimento a questa eventualità, né tantomeno alle soluzioni individuate per farvi fronte[106].

Qualunque sia il metodo utilizzato per la determinazione dei valori percentili a cui far corrispondere le soglie, la distanza tra i valori-soglia che definiscono ciascuna classe varia a seconda della forma della distribuzione; da un lato ciò permette la neutralizzazione delle differenze di campi di variazione, medie e varianze, dall'altro non assicura che le classi siano omogenee al loro interno ed eterogenee tra loro. In altri termini in una distribuzione asimmetrica e leptocurtica, come in genere sono le distribuzioni bibliometriche, è ragionevolmente possibile immaginare che scegliendo la soglia al 50% della distribuzione tutti i casi al di sotto della soglia presentino valori molto bassi e molto diversi, ad esempio, da quelli dei casi al di sopra della soglia che definisce il 20% più alto della distribuzione. Se invece in uno specifico campo di studi la funzione di distribuzione cumulativa dovesse risultare platicutrica i valori presentati dai casi nella classe più alta non sarebbero molto diversi da quelli presentati dai casi nella classe più bassa. E' chiaro come questo approccio sia necessario in relazione all'obiettivo valutativo della procedura: le soglie non sono definite a priori proprio in ragione dell'elevata varietà delle distribuzioni, nulla da dire dal punto di vista tecnico. Nondimeno da un punto di vista strettamente semantico l'arbitrarietà delle percentuali cumulate che fungono da soglie può condurre a una classificazione anche molto diversa di articoli che presentino differenze minime sul valore effettivo dell'indicatore.

Un'ulteriore questione da affrontare è quella relativa alla scelta della *subject category (SC)* o categoria dell'*all science journal classification (ASJC)*, in cui considerare il prodotto. Le riviste infatti, tanto in WoS quanto in Scopus, possono appartenere a più di una categoria. Nell'Area 3 in tutti i caso in cui il prodotto risultasse classificato in più di una SC (o ASJC) è stata presa in considerazione l'indicazione del soggetto valutato autore del prodotto da valutare, modificata dal GEV ove necessario (Anvur, 2013d, GEV3, p. 90)[107].

In entrambi i database inoltre è prevista una categoria per la produzione scientifica multidisciplinare (*multidisciplinary science*), cui appartengono riviste come *Science* o *Nature*, che ospitano articoli riguardanti una pluralità di argomenti. Tutti gli articoli pubblicati su riviste appartenenti a questa categoria sono stati riassegnati in base alle citazioni contenute nell'articolo, cioè sulla base delle riviste citate si individuerà la *subject category (SC)* o la classe dell'*all science*

[106] Le interviste ai membri del GEV e del Consiglio Ditettivo Anvur non hanno permesso di chiarire questo passaggio, ma il dottor Anfossi (assistente GEV di Area 2 nel corso della VQR, addetto tra l'altro alla progettazione del sistema informatico per la valutazione dei prodotti) ha chiarito che la soluzione adottata prevedeva sempre l'allargamento della soglia più alta. E' chiaro che in casi estremi questa scelta avrebbe potuto produrre classificazioni prive dell'ultima classe, tuttavia appare corretta nei confronti dei soggetti valutati e sostanzialmente in linea con l'impostazione generale per cui in caso di valutazioni discordanti viene utilizzata la più favorevole.

[107] Nel corso delle interviste sia il professor Barone che il professor Pacchioni hanno sostenuto che la categoria tematica utilizzata per la valutazione era la più favorevole al prodotto, ma queste affermazioni sembrano derivare da un effetto memoria, dato che questo criterio era usato spesso per dirimere le controversie. Il dottor Anfossi ha però escluso questa possibilità e confermato che, a meno di modifiche da parte del GEV effettuate sui singoli prodotti, l'algoritmo era predisposto per valutare ciascun prodotto nella classe di merito segnalata da chi lo aveva sottomesso, come esposto nel report finale di Area. Le parole di Bonaccorsi confermano questa evidenza: «naturalmente c'è un problema legato alle pubblicazioni multidisciplinari, così dette, per le quali era stato immaginato anche un percorso di peer review complementare all'analisi bibliometrica, però in ogni caso era il soggetto a selezionare la SC in cui essere valutato, per cui da questo punto di vista credo che problemi di qualità del dato non ci fossero» (Intervista Bonaccorsi).

journal classification (ASJC). L'articolo porterà con sé nella nuova categoria l'*impact factor* della rivista e il numero di citazioni ricevute.

Uno degli elementi meno chiari circa la procedura di valutazione bibliometrica diretta è la successione temporale tra il calcolo delle soglie (tanto per gli indici di impatto quanto per le citazioni) e la scelta della categoria. Nel caso in cui le soglie vengano calcolate prima dell'assegnazione della categoria, infatti, c'è il rischio (almeno per le riviste classificate in *multidisciplinary science*) che la rivista o l'articolo valutato non contribuisca a creare la distribuzione, dunque le soglie utilizzate per la valutazione. Tra l'altro questa sequenza creerebbe una mancanza di rispondenza dell'ampiezza delle classi alla loro definizione[108].

Dal punto di vista procedurale, ricondurre gli articoli pubblicati in riviste classificate come multidisciplinari a una specifica categoria prima di riportare gli indicatori in classi di merito sarebbe la scelta più corretta. Il confronto infatti in bibliometria va, per quanto possibile, sempre effettuato tra elementi simili. In questo caso, volendo valutare ad esempio un articolo di chimica analitica comparso su *Science* sarebbe stato scorretto confrontarlo con altri articoli, sempre pubblicati in riviste multidisciplinari, ma riguardanti discipline completamente diverse: biotecnologia, immunologia, astrofisica, fino a storia contemporanea o antropologia culturale.

Il fatto che gli articoli pubblicati in riviste multidisciplinari presentino un numero di citazioni mediamente più elevato (Gòmez Nuñez *et al.* 2011), non dovrebbe introdurre distorsioni nella classificazione. L'introduzione nelle distribuzioni di valori più elevati riflette l'effettivo potenziale citazionale del campo di studi, che non è limitato alle riviste di settore, ma include le riviste multidisciplinari. Dal punto di vista della classificazione dei singoli prodotti ci si può attendere che gli articoli provenienti da riviste multidisciplinari abbiano una maggiore probabilità di essere classificati come eccellenti, ma ciò risulta corretto e coerente con gli obiettivi classificatori dato che riflette il loro maggior numero di citazioni e la loro collocazione in riviste ad alto impatto.

Le differenze tra le classificazioni di WoS e Scopus non sono tra le questioni più discusse nella letteratura bibliometrica, anche se con alcune eccezioni, legate a specifici campi di studio (si veda ad esempio Lopez-Illescas *et al.* 2008; Abrizah *et al.* 2013). Non è da escludere che queste differenze siano in grado di produrre valutazioni non concordanti di uno stesso prodotto, dato il fatto che una stessa rivista può conseguire posizioni diverse nei ranking a seconda della categoria tematica (Amin e Mabe, 2000) e che i criteri su cui queste classificazioni si fondano nonché la qualità del risultato sono tutt'altro che indiscussi (Boyack *et al.* 2005; Leydesdorff e Rafols, 2009; Jacsó, 2013; Gómez-Núñez *et al.* 2011 e 2014). Tale questione dipende direttamente dalla struttura dei database (è infatti approfondita nel § 5.2.1), per ora è sufficiente sottolineare che queste classificazioni non sono sovrapponibili né è chiaro se una stessa rivista sia classificata in categorie semanticamente affini nei due database. Data la procedura di riconduzione degli indicatori alle classi di merito è del tutto plausibile che la classe in cui il prodotto viene considerato sia in grado di influire significativamente sulla sua valutazione, non solo considerando classi simili nei due database, ma anche considerando classi differenti nello stesso database.

[108] Di nuovo le interviste non sono state utili alla precisazione di questi passaggi, il dottor Anfossi tuttavia ha chiarito che l'algoritmo calcolava le distribuzioni cumulate all'interno di ciascuna categoria e poi confrontava i valori relativi agli indicatori dei prodotti classificati come multidisciplinari con le soglie già calcolate per la categoria di destinazione.

Infine vale la pena sottolineare che tanto le *subject categories* quanto le categorie dell'ASJC classificano riviste, non singoli articoli, mentre l'obiettivo della VQR è la valutazione dei singoli prodotti. Non è possibile escludere che un articolo pubblicato in una rivista di chimica-fisica possa in effetti trattare questioni legate all'elettrochimica o ai nano materiali, dunque le *subject categories* non sono in grado di assicurare che la procedura confronti i prodotti all'interno di campi disciplinari omogenei. La classificazione dei singoli articoli per campi disciplinari omogenei non è impossibile (si veda ad esempio la procedura proposta in Waltman e van Eck, 2012), tuttavia richiederebbe un enorme investimento di tempo e risorse, che in una prospettiva economica può giustificare l'utilizzo di categorie pre-costituite (*cfr*. § 6.1.2).

4.3 La sintesi e le classi di merito

La sintesi dei dati rilevati è l'ultimo passo del modello lazarsfeldiano per l'operativizzazione dei concetti, e prevede la ricomposizione degli indicatori in indici, cioè in misure sintetiche, delle dimensioni del concetto e/o del concetto stesso[109].

Nobile (2008) segnala come questa fase della procedura lazarfeldiana sia la più trascurata nel dibattito metodologico. In ragione del tipo di variabili da sintetizzare è possibile individuare due vie alla costruzione degli indici: una via analitica, utilizzabile nel caso in cui le variabili siano categoriali, e una via matematica, servibile esclusivamente nel caso in cui si abbia a che fare con variabili cardinali o quasi-cardinali (Marradi, 1984, 2007; Nobile, 2008). Mentre il numero di procedure riconducibili all'alveo della sintesi analitica è abbastanza ridotto, le procedure di sintesi per via matematica sono numerose ed estremamente diversificate tra loro in ragione della logica sottostante e della complessità delle operazioni da effettuare.

La via analitica alla costruzione di indici sintetici passa essenzialmente attraverso la riduzione di uno spazio di attributi. Marradi (2007) pone l'accento su due questioni essenziali nella costruzione degli indici tipologici: innanzitutto si deve tener conto della dimensione semantica, e in secondo luogo della dimensione numerica delle classi dell'indice finale. Si tratta di seguire gli stessi criteri della classificazione mirando a raggiungere la massima omogeneità tra stati inclusi nella stessa categoria e la massima eterogeneità tra stati inclusi in categorie diverse, evitando squilibri nella numerosità delle categorie stesse. Naturalmente «occorre bilanciare il criterio semantico con quello numerico» (*ibidem*, p. 185). Nobile (2008) presenta una serie di questioni centrali, riconducibili proprio al principio dell'adeguatezza dell'indice ottenuto al suo scopo operativo e al suo significato semantico:

a) tenere conto della effettiva distribuzione dei dati per evitare che la distribuzione dell'indice ottenuto risulti sbilanciata o non in grado di rilevare le differenze nel campione d'indagine;
b) tenere conto della dimensione semantica delle due variabili nella costruzione dell'indice: se una delle due dovesse avere una parte indicante maggiore rispetto a quella dell'altra dovrebbe assumere un peso maggiore;

[109] L'indice è l'esito finale del processo di operativizzazione, dunque il termine è riferibile tanto agli indici sintetici quanto alle variabili in grado di rappresentare da sole concetti (strutturalmente) semplici. In questa sede l'attenzione di focalizza sulla sintesi dell'informazione apportata da più indicatori in un unico indice.

c) fare attenzione alla simmetria-asimmetria dello spazio di attributi (cioè ai differenti casi che si hanno nel caso in cui le variabili abbiano o meno lo stesso numero di modalità) sia in riferimento all'aspetto semantico che in riferimento alla distribuzione effettiva dei dati;
d) tenere conto dell'obiettivo dell'indice nella decisione relativa al numero di classi da costruire.

La costruzione di un indice tipologico a partire da due variabili categoriali non è che il caso più semplice, vi sono alcune accortezze in grado di ridurre le difficoltà nei casi più complessi (Nobile, 2008), ma non è possibile eliminarle, infatti «ogni volta che si vogliono combinare uno o più indicatori, la complessità semantica aumenta in modo esponenziale, le decisioni necessarie per ridurre la tipologia sono sempre più problematiche e il controllo intellettuale sull'intera operazione e sui suoi risultati diminuisce» (Marradi, 2007, p. 186). Aumentando il numero degli indicatori da sintetizzare si indebolisce il controllo sul significato delle combinazioni oppure si rende necessario aumentare il numero delle categorie dell'indice finale, «per questo motivo ogni volta che si possono immaginare definizioni operative che creano variabili cardinali o quasi cardinali, i ricercatori lo preferiscono» (*ibidem*).

La via matematica alla costruzione di indici sintetici offre una varietà di opzioni: tecniche additive, tecniche di combinazione[110], tecniche fattoriali[111], di scaling o sociometriche[112] (Nobile, 2008)[113].

Le tecniche additive sono probabilmente le più diffuse, anche in ragione della loro apparente semplicità, e sono essenzialmente basate sulla somma delle variabili che compongono l'indice[114]. Naturalmente questo insieme di tecniche è pienamente legittimo solo nel caso in cui si operi con variabili cardinali o quasi-cardinali, tuttavia è possibile sfruttarli da un punto di vista procedurale anche in riferimento a variabili categoriali. Infatti, è possibile ricorrere agli indici per somma anche

[110] Le tecniche di combinazione, che si riferiscono alla relazione tra due valori, includono tassi, rapporti, quozienti, prevedono spesso un doppio livello di sintesi, confrontando misure già sintetiche, e si ottengono con procedure semplici e consolidate, dunque in genere non pongono particolari problemi dal punto di vista metodologico.

[111] Le tecniche fattoriali sono invece basate sul risultato di procedure di analisi multivariata, come ad esempio l'analisi fattoriale, in grado di sintetizzare gran parte dell'informazione portata da una serie di indicatori cardinali o quasi cardinali in pochi fattori. Questo genere di indici necessita di una attenta analisi semantica, ma dal punto di vista prettamente procedurale garantiscono un elevatissimo livello di standardizzazione.

[112] Svariate procedure di sintesi sono strettamente connesse alle tecniche di rilevazione dei dati, e a specifici ambiti teorici, generalmente prestabilite e largamente consolidate. Basti pensare alle tecniche di scaling, per le quali i punteggi sull'indice vengono attribuiti ai casi con procedure consolidate e note, o alle tecniche sociometriche, che producono misure di sintesi sulla base di procedure univocamente definite e allo stesso tempo semplici (in genere combinazioni di valori) che possono essere attribuite ai singoli casi della matrice.

[113] Alle classi presentate, Nobile aggiunge la classe residuale delle tecniche *idiomatiche*, cioè l'insieme delle tecniche di sintesi che, pur non essendo riconducibili a nessuna delle classi precedenti, ricorrono comunque ad algoritmi di natura matematica.

[114] In riferimento agli indici additivi Marradi (1984) sottolinea la necessità di quattro condizioni:

- *condizione fattuale*: non devono mancare dati su uno o più indicatori (una soluzione è la standardizzazione);
- *condizione numerica*: tutte le variabili che si sommano devono avere la stessa estensione di scala (una soluzione è la normalizzazione);
- prima *condizione semantica*: la direzione semantica delle variabili sommate deve essere la stessa (la direzione semantica può essere invertita dal ricercatore ove lo ritenga opportuno);
- seconda *condizione semantica*: gli indicatori da sommare devono essere ritenuti ugualmente validi (cioè nella stessa relazione semantica con il concetto) o il ricercatore potrebbe optare per una ponderazione sulla base della validità che gli attribuisce su base semantica o in seguito a una analisi fattoriale.

«in un'ottica interamente stipulativa, in cui la combinazione dei punteggi assegnati a due variabili categoriali fa scaturire una variabile indice alla quale viene dato un ulteriore punteggio stipulativo»; in questo caso «un indice per somma non va realmente a sommare i punteggi, ma serve soltanto a passare rapidamente all'attribuzione dei valori-indice senza dover fare ricorso a operatori logici» (Nobile, 2008, p. 70). Lo stesso discorso può essere applicato alle scale con categorie ordinate, ma in questo caso è necessario tenere conto del fatto che utilizzare una procedura additiva significa «forzare la natura delle variabili, portandole a un livello quasi cardinale» (*ibidem*, p. 72).

La fase finale del processo di operativizzazione richiede una serie di decisioni tecniche, semantiche e sintattiche ognuna delle quali può avere delle ripercussioni sull'attendibilità del dato finale nel caso una o più parti della procedura siano errate, o anche sulla sua validità nel caso in cui uno o più indicatori sintetizzati presentino problematiche da questo punto di vista.

Nell'ambito della ricerca sociale i dati sono infatti l'esito di un processo di progettazione e costruzione che include una serie di operazioni logico-semantiche e tecnico-operative. La qualità del dato dipende in grandissima parte dalle fasi di concettualizzazione e scomposizione dei concetti, la sua validità può essere ricondotta alla fase relativa alla selezione degli indici, la sua affidabilità/attendibilità alle fasi della rilevazione e del trattamento statistico dei dati, ma qualsiasi problema sia stato trascurato nelle fasi precedenti si ripercuote e può essere amplificato nella fase della costruzione degli indici sintetici (Nobile, 2008).

4.3.1 La sintesi e le classi di merito nella VQR

La sintesi dei dati nella VQR presenta delle differenze in ragione dell'approccio valutativo bibliometrico o meno. Nel primo caso le informazioni fornite dai due indicatori vengono sintetizzate direttamente con l'assegnazione al prodotto di una classe di merito, nel caso della valutazione tramite peer review i passaggi sono due: innanzitutto è necessario sintetizzare i punteggi assegnati dai singoli revisori su ciascun criterio (rilevanza, originalità e internazionalizzazione) assegnando al prodotto una classe di merito, in seguito le classi assegnate al prodotto da due (o più) revisori devono essere confrontate e sintetizzate nella classe di merito finale.

L'assegnazione della classe di merito finale a ciascun prodotto, tanto per la peer review che per la procedura bibliometrica, avveniva nell'ambito dei GEV. A questo scopo «ogni GEV ha costituito dei gruppi di consenso formati da due o tre membri, che, sulla base dei punteggi espressi dai due (o più) revisori e di un procedimento definito a priori, perveniva alla classificazione finale. […] Al di là di piccole variazioni tra i GEV, il procedimento richiedeva l'approvazione del gruppo di consenso nel casi di valutazioni *peer* coincidenti o con differenze di una sola classe, mentre consentiva la richiesta di una terza revisione *peer* nel caso di valutazioni discordanti per due o tre classi» (Anvur, 2013a, p. 26). Quanto riportato all'interno dei report di Area o delle relative appendici va più o meno nella stessa direzione, ad esempio per l'Area 14 «nel caso di valutazioni non convergenti dei revisori *peer*, il sub-GEV creerà 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*. Il Gruppo di Consenso potrà avvalersi anche del giudizio di un terzo esperto. In ogni caso la responsabilità della valutazione conclusiva è dell'intero GEV» (Anvur, 2013d, GEV 14, Appendice, p. 74).

In riferimento alla peer review nei report finali sono disponibili pochissime informazioni puntuali circa l'assegnazione della classe di merito al prodotto sulla base dei punteggi ottenuti nelle schede di valutazione. Nel rapporto finale si legge che: «la somma dei tre punteggi era confrontata con tre soglie per generare una classificazione finale in quattro classi. La classificazione era proposta al revisore per consentirgli di confrontarla con la definizione delle classi 1, 2, 3 e 4 della Sezione 2.5 e, eventualmente, di modificare i punteggi» (Anvur, 2013a, p. 26). Sono già state segnalate alcune controindicazioni legate al "controllo" diretto da parte del revisore della classe di merito assegnata al prodotto (*cfr*. § 4.1.2), qui è interessante notare che le informazioni fornite nei rapporti risultano ancora una volta generiche e superficiali. Ad esempio, nel documento relativo ai criteri di valutazione del GEV 14 si legge soltanto che: «il GEV trasformerà le indicazioni contenute nella scheda revisore in una delle 4 classi finali di merito» (Anvur, 2013d, GEV 14, Appendice, p. 74). L'unica eccezione è rappresentata dall'Area 13 delle Scienze Economiche e Statistiche che, presentando la propria scheda di valutazione, dà conto anche dei punteggi e dell'assegnazione delle classi di merito (Anvur, 2013d, Appendice, p. 113).

Le informazioni circa questo passaggio sono però reperibili nel documento circa i "Criteri di assegnazione della classi di merito nel caso di valutazioni peer review con valutazioni non coincidenti da parte dei *referee*" (Anvur, 2014), che presenta brevemente le proposte di classificazione del caso di due, tre, quattro o cinque referaggi. Sostanzialmente viene presentata «la corrispondenza tra la somma dei punteggi attributi dai *referee* e le classi di merito da utilizzare nel caso di giudizi non concordanti tra *referee* e nei casi in cui la classe di merito assegnata dai *referee* non si discosti per più di una classe di merito» (*ibidem*, p. 1). Le proposte sono distinte sulla base della scala di valutazione prevista nella scheda per ciascuna Area, cioè per le scale da 1 a 9 punti (Aree: 2, 3, 6, 7, 8, 10, 11, 12, 13, in Tabella 18) e per quelle da 0 a 3 (Aree: 1, 4, 5, 9, 14, in Tabella 19).

Tabella 18 - Proposte di classificazione per due, tre, quattro o cinque referaggi per le Aree con punteggi in scala 1-9 (adattamento da Anvur, 2014, p. 2)

Classe	Punteggio	S1+S2	S1+S2+S3	S1+S2+S3+S4	S1+S2+S3+S4+S5
Eccellente	23-27	46-54	69-81	92-108	115-135
Buono	18-22	36-45	54-68	72-91	90-144
Accettabile	15-17	30-35	45-53	60-71	75-89
Limitato	3-14	6-29	9-44	9-45	15-74

Tabella 19 - Proposte di classificazione per due, tre, quattro o cinque referaggi per le Aree con punteggi in scala 0-3 (adattamento da Anvur, 2014, p. 2)

Classe	Punteggio	S1+S2	S1+S2+S3	S1+S2+S3+S4	S1+S2+S3+S4+S5
Eccellente	8-9	16-18	24-27	32-36	40-45
Buono	6-7	12-15	18-23	24-31	30-39
Accettabile	5	10-11	15-17	20-23	25-29
Limitato	0-4	0-9	0-14	0-19	0-24

Purtroppo il documento non contiene alcuna argomentazione a supporto della selezione delle soglie o della scelta di sintetizzare i punteggi per via sommatoria piuttosto che sintetizzare direttamente le classi per via tipologica. Si intuisce la volontà di rispecchiare, almeno in parte, la definizione quantitativa delle classi di merito, con il riferimento a quote percentuali di punteggi

corrispondenti all'incirca al 20% per l'eccellenza, al 20% per buono al 10% per accettabile e al 50% per limitato, pur con i limiti connessi al numero di esiti possibili[115].

La classificazione dei prodotti nelle classi di merito finali viene trattata più estesamente in riferimento all'analisi bilbiometrica, in quasi tutti i GEV[116] la procedura prevedeva «la costruzione di una matrice 4x4 i cui elementi sono individuati dai valori della coppia di indicatori» sulla cui base «l'assegnazione di una fra cinque classi finali a ciascuna delle 16 celle: le 4 classi di merito della VQR e una classe *undecided* (IR, che sta per "*informed review*") caratterizzata da indicazioni divergenti dei due indicatori; gli articoli che ricadevano in quest'ultima classe sono stati valutati utilizzando la *informed* peer review*»* (Anvur, 2013a, p. 23).

Le matrici utilizzate per l'attribuzione delle classi di merito finali sono state diverse da GEV a GEV, ma anche all'interno dei singoli GEV le matrici utilizzate sono state diverse in base ad esempio al settore disciplinare[117] o all'anno di pubblicazione del prodotto[118].

Va tenuto conto del fatto che l'obiettivo finale della valutazione è l'assegnazione di ciascun prodotto a una *classe di merito* nella «scala di valore condivisa a livello internazionale» (Anvur, 2011a, p. 7). Nonostante l'attenzione a questo aspetto, le quote, ottenute applicando gli algoritmi utilizzati nel corso della VQR all'intera produzione scientifica mondiale, non corrispondono alla distribuzione teorica. In sostanza la probabilità che un prodotto venga classificato come eccellente risulta più alta del previsto, mentre la probabilità che un prodotto sia valutato come limitato è ristretta (Anvur, 2011a, Appendice A), per questa ragione nel rapporto finale dell'Anvur un'intera appendice è dedicata alla "calibrazione" degli algoritmi bibliometrici.

La mancata calibratura degli algoritmi conduce a una classificazione non rispondente ai criteri stabiliti dal decreto ministeriale e dal bando, ed è in grado almeno in parte di spiegare perché tutte le Aree prevalentemente bibliometriche presentino una quota di prodotti eccellenti più elevata rispetto alle altre (*cfr.* Grafico 2, p. 71).

4.3.2 La sintesi e le classi di merito nell'Area delle Scienze Politiche e Sociali

La sintesi dei giudizi nell'Area delle Scienze Politiche e Sociali prevedeva due passaggi: nel primo a partire dai giudizi espressi per ciascun criterio da un singolo revisore era necessario ottenere una classe di merito a cui assegnare il prodotto, nel secondo le due classi di merito ottenute dal prodotto a partire dal giudizio dei due revisori dovevano essere sintetizzate nella classe di merito finale.

[115] Bonaccorsi a questo proposito ha spiegato come: «fondamentalmente dovevamo avere la mappatura tra il punteggio e la scala imposta dal bando» (Intervista Bonaccorsi), e questa ipotesi trova riscontro anche nelle parole del Presidente GEV di Area 14: «praticamente non abbiamo avuto moltissima libertà nel decidere le soglie, nel senso che adesso non voglio dire cose che non ricordo esattamente, non mi ricordo se l'abbiamo deciso in sede di presidenza, però se non l'abbiamo deciso in sede di conferenza dei presidenti è stata una scelta tecnica fatta dagli assistenti insieme a Sergio Benedetto, il coordinatore generale, per cui di fatto non c'è stata un'autonomia neanche su questo» (Intervista Colozzi).

[116] Sono già state sottolineate le particolarità della procedura utilizzata dal GEV 13 di Scienze Economiche e Statistiche (§ 4.1.1).

[117] E' il caso del GEV 9, che utilizza una procedura differente per il SSD INGINF/05 (Anvur, 2013d, GEV 9, Appendice, pp. 65).

[118] E' il caso, ad esempio del GEV 6 (Anvur, 2013d, GEV 6, Appendice, pp. 22-24).

Sul primo passaggio non sono disponibili informazioni nel report finale di Area o nelle sue appendici, dunque solo grazie al documento relativo ai criteri di assegnazione delle classi (Anvur, 2014), è stato possibile descrivere e analizzare questa fase della procedura. Si è già evidenziato che a ciascuna delle modalità di risposta previste per le tre domande della scheda di valutazione corrispondeva un punteggio, espresso con un numero naturale; la sintesi di questi tre punteggi prevedeva in primo luogo una somma semplice:

$$P_{tot} = P_{D1} + P_{D2} + P_{D3}$$

Il risultato di questa operazione ha un campo di variazione incluso tra 0 (0+0+0, il punteggio minimo su tutti e tre i criteri) e 9 (3+3+3, il punteggio massimo su tutti e tre i criteri).

La scelta di sintetizzare le tre variabili con una somma semplice implica dal punto di vista concettuale l'assunzione di un uguale rapporto semantico di ciascuno dei tre indicatori con il concetto di qualità della ricerca. Si è già sottolineato quanto questo assunto sia sostenibile almeno quanto l'assunto contrario, ma che proprio per questo andrebbe esplicitato e argomentato in nome di quei criteri di *pubblicità*, *replicabilità* e *controllabilità* a cui così spesso si è fatto riferimento nel corso della trattazione.

L'esito di questa somma veniva confrontato con delle soglie e ricondotto a una delle quattro classi di merito, come indicato nel report finale dell'Agenzia: «la somma dei tre punteggi era confrontata con tre soglie per generare una classificazione finale in quattro classi» (Anvur, 2013a, p. 26). Lo schema di riconduzione presentato dall'Anvur sui criteri di attribuzione delle classi di merito (Anvur, 2014) è riportato in Tabella 20, si noti che corrisponde perfettamente alle quote previste dal bando per le classi di merito. Assumendo cioè che la «scala di valore condivisa a livello internazionale» (Anvur, 2011a, p. 7) abbia un campo di variazione da 0 a 9, nel rispetto del criterio quantitativo le classi di merito andrebbero assegnate esattamente come nello schema.

Tabella 20 - Corrispondenza tra i punteggi totali e le classi di merito per l'Area 14

Classe	Punteggi	Quota
Eccellente	8 - 9	20%
Buono	7 - 6	20%
Accettabile	5	10%
Limitato	4 - 3 - 2 - 1 - 0	50%

Una prima fondamentale osservazione da avanzare è che l'attribuzione delle classi non tiene in alcun conto la distribuzione empirica dei punteggi. Ciò significa che ogni prodotto viene classificato indipendentemente da tutti gli altri, se ad esempio nessuno dei prodotti dell'Area fosse risultato *molto originale e innovativo* nessun prodotto sarebbe stato classificato come eccellente. Questa scelta indica la volontà di valutare i prodotti in base alle loro caratteristiche, senza tener conto delle caratteristiche di tutti gli altri prodotti scientifici, dunque in base a standard predefiniti. Una soluzione alternativa avrebbe potuto assegnare i punteggi alle classi tenendo conto della loro distribuzione empirica, considerando così le classi di merito come *relative* all'insieme dei prodotti presentati. E' in relazione agli standard predefiniti che la scelta di permettere ai revisori di modificare i punteggi assegnati ai prodotti pone dei problemi: gli standard del revisore potrebbero infatti risultare diversi da quelli stabiliti ex-ante per l'esercizio di valutazione, e dunque la nuova classe di merito potrebbe essere soggetta a distorsioni non controllabili.

Le tre variabili assumono lo stesso peso nella costruzione dell'indice finale. Considerando ciascun punteggio come corrispondente a una classe di merito (3 a eccellente, 2 a buono, 1 ad accettabile e 0 a limitato), dal punto di vista semantico le soglie sono traducibili nei seguenti asserti:

- un prodotto è *eccellente* se risulta eccellente su tutti e tre i criteri, oppure eccellente su almeno due criteri e buono sul terzo;
- un prodotto è *buono* se risulta buono su tutti e tre i criteri, oppure eccellente su almeno due criteri, pur essendo accettabile o limitato sul terzo, oppure ancora se è eccellente su un criterio e buono sugli altri due;
- un prodotto è *accettabile* se risulta accettabile su tutti e tre i criteri, o eccellente su un criterio, buono sul secondo ma limitato sul terzo, oppure buono su due criteri e accettabile sul terzo;
- un prodotto è *limitato* se risulta limitato su tutti e tre i criteri, oppure limitato o accettabile su almeno due dei tre criteri, pur essendo eccellente sul terzo, oppure se è buono su due criteri e limitato sul terzo.

La somma qui «serve soltanto a passare rapidamente all'attribuzione dei valori-indice senza dover fare ricorso a operatori logici» (Nobile, 2008, p. 70). Allo scopo di sottolineare il criterio semantico alla base della determinazione delle soglie si riporta una rappresentazione tipologica dell'assegnazione delle classi di merito in base ai punteggi (Tabella 21), ed il numero assoluto e relativo di combinazioni che danno luogo all'assegnazione di ciascuna classe (Tabella 22).

Tabella 21 – Rappresentazione tipologica dell'attribuzione delle classi di merito nell'Area 14

D1	D2	D3			
		0	1	2	3
0	0	L	L	L	L
	1	L	L	L	L
	2	L	L	L	A
	3	L	L	A	B
1	0	L	L	L	L
	1	L	L	L	A
	2	L	L	A	B
	3	L	A	B	B
2	0	L	L	L	A
	1	L	L	A	B
	2	L	A	B	B
	3	A	B	B	E
3	0	L	L	A	B
	1	L	A	B	B
	2	A	B	B	E
	3	B	B	E	E

Tabella 22 - Corrispondenza tra i punteggi totali e le classi di merito per l'Area 14, con indicazione del numero assoluto e relativo di combinazioni di punteggi che conducono all'assegnazione di ciascuna classe

Classe	Punteggi	Combinazioni	
Eccellente	8 - 9	4	6,25%
Buono	7 - 6	16	25%
Accettabile	5	12	18,75%
Limitato	4 - 3 - 2 - 1 - 0	32	50%

Il fatto che la determinazione delle soglie appaia sostanzialmente fondata su considerazioni di ordine semantico prima che di ordine matematico non esonera il GEV né l'Agenzia dall'obbligo di

esplicitarle e giustificarle[119]. Infatti la scelta delle soglie, pur rispettando la definizione quantitativa delle classi, ha delle implicazioni semantiche che non riflettono precisamente la definizione delle classi di merito fornita dall'Anvur (Anvur, 2011a, p. 7). Si è già discusso di queste definizioni e delle relative problematiche di ambiguità e vaghezza (*cfr*. § 3.2), ma è importante considerare le soglie alla luce della definizione semantica delle classi di merito, considerando che la costruzione di un indice sintetico dovrebbe riflettere la definizione del concetto cui si riferisce. Confrontando le definizioni riportate dall'Anvur con quelle effettivamente deducibili dalla scelta delle soglie è possibile notare non poche differenze, alcune più rilevanti di altre. In sostanza mentre nelle definizioni i criteri di originalità e rilevanza sembrano legati da un legame di co-implicazione e l'internazionalità non assume un ruolo decisivo nel determinare l'assegnazione della classe di merito, la scelta delle soglie sottintende non solo che ciascun criterio assuma lo stesso peso nella determinazione della classe di merito, ma anche che i criteri possano essere ortogonali tra loro.

In relazione al primo punto, nella sintesi dei giudizi l'internazionalizzazione pesa quanto gli due criteri, mentre nella definizione Anvur di tutte le classi di merito (eccellente, buono, accettabile, limitato) si fa riferimento tanto al livello nazionale quanto al livello internazionale.

In relazione al secondo punto la scelta delle soglie non riflette le relazioni tra le dimensioni concettuali né il legame tra le dimensioni e il concetto di qualità della ricerca. Nelle definizioni non vi è alcun riferimento a tutta una serie di casi, quelli in cui la classificazione sui tre criteri risulti discordante (ad esempio il caso in cui un prodotto risulti eccellente sul piano della rilevanza, ma limitato sul piano della originalità e dell'internazionalizzazione), mentre la definizione delle soglie permette di ricondurre ciascuno di questi casi a una delle classi di merito, anche in assenza di riferimenti chiari nella loro definizione (*cfr*. § 3.2).

In particolare con riferimento al livello *limitato* la soglia scelta e le sue implicazioni semantiche non riflettono la definizione, includendo non solo i prodotti che ottengono il punteggio minimo su rilevanza e originalità, ma anche prodotti che, ad esempio, ottengono un punteggio medio-basso (corrispondente in sostanza al livello accettabile) su questi due criteri ma il minimo sull'internazionalizzazione, oppure il minimo su internazionalizzazione e originalità e il massimo sul criterio della rilevanza. Ancora una volta la mancanza di chiarezza nella definizione del concetto si riflette nella sintesi dei giudizi, generando incongruenze e mettendo a rischio la validità del dato sintetico.

La sintesi dei giudizi di un singolo revisore su un singolo prodotto non è che il primo passo verso l'attribuzione della classe di merito finale. Il report finale segnala che, in riferimento a ciascun prodotto, un gruppo di consenso, formato da due o tre membri del GEV, «sulla base dei punteggi espressi dai due (o più) revisori e di un procedimento definito a priori, perveniva alla classificazione finale. […] Al di là di piccole variazioni tra i GEV, il procedimento richiedeva l'approvazione del gruppo di consenso nel casi di valutazioni *peer* coincidenti o con differenze di una sola classe, mentre consentiva la richiesta di una terza revisione *peer* nel caso di valutazioni discordanti per due o tre classi» (Anvur, 2013a, p. 26).

[119] Il presidente GEV segnala anche alcuni rischi derivanti dal numero di combinazioni riconducibili all'eccellenza e le caratteristiche della produzione scientifica dell'Area: «siccome l'eccellenza era 9, noi di 9 non ne abbiamo, se non i pochissimi casi di lavori pubblicati su riviste internazionali e valutati di classe A, cioè eccellenti» (Intervista Colozzi).

Sembra però che i gruppi di consenso abbiano proceduto all'assegnazione della classe di merito finale solo per le valutazioni divergenti di più di una classe, dato che soglie pre-stabilite erano previste anche per la corrispondenza tra i punteggi assegnati dai revisori e la classe di merito finale[120]. Nell'Appendice A si legge infatti che «il GEV trasformerà le indicazioni contenute nella scheda revisore in una delle 4 classi finali di merito», e poco oltre che: «nel caso di valutazioni non convergenti dei revisori *peer*, il sub-GEV creerà 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*. Il Gruppo di Consenso potrà avvalersi anche del giudizio di un terzo esperto» (Anvur, 2013d, GEV 14, Appendice, p.74).

Il rapporto e le appendici non contengono informazioni sulle modalità di costituzione e gestione dei gruppi di consenso, tuttavia è emerso dalle interviste che i gruppi di consenso erano formati dai due EV che avevano in carico il prodotto, cui si aggiungevano un terzo e più raramente un quarto membro, in genere il Presidente del GEV, il vicepresidente oppure il coordinatore sub-GEV di riferimento[121].

Riguardo le modalità di discussione e valutazione interne ai gruppi di consenso, nel documento sui criteri di valutazione dell'Area 14, sono presenti solo alcune indicazioni: «in ogni caso, la responsabilità della valutazione conclusiva è dell'intero GEV, che terrà conto delle valutazioni *peer* e, anche, delle caratteristiche della collana editoriale nella quale la monografia è stata pubblicata: si valuterà l'esistenza di un comitato editoriale, di procedure trasparenti di revisione per decidere sulla pubblicazione, della diffusione a livello nazionale e internazionale dei prodotti dell'editore, di recensioni dell'opera pubblicate su riviste internazionali, e ogni altro elemento atto a fornire indicazioni utili sulla qualità e impatto dell'opera» (Anvur, 2013d, GEV 14, Appendice, p. 74).

E' da sottolineare che nell'Area delle Scienze Politiche e Sociali la procedura utilizzata è stata quella della semplice peer review, non dell'*informed* peer review*,* nonostante quanto riportato nel documento relativo ai criteri (*ibidem*)[122]. In teoria la procedura di valutazione degli articoli avrebbe dovuto basarsi non solo delle dei giudizi dei pari, ma anche : «per gli articoli pubblicati su riviste indicizzate in WoS e/o Scopus, sui dati bibliometrici derivanti da un algoritmo che tiene conto, in misura diversa a seconda della data di pubblicazione dell'articolo, sia del numero di citazioni che dell'indicatore bibliometrico della rivista ospitante. Per gli articoli apparsi su riviste italiane, il GEV ha formulato una classificazione delle riviste» (*ibidem*). La classificazione delle riviste e le altre

[120] Infatti la procedura non prevedeva necessariamente la discussione da parte del GEV: «noi abbiamo fissato delle soglie che permettevano di andare in automatico al numero massimo di prodotti. Abbiamo discusso direttamente quelli che avevano problemi e che quindi non rientravano chiaramente in una soglia o nell'altra soglia. Solo su quelli» (Intervista Colozzi).

[121] Nelle parole del Presidente del GEV: «i gruppi di consenso sono stati formati usando fondamentalmente il criterio iniziale di assegnazione dei referaggi […] Lo stesso criterio è stato utilizzato per i gruppi di consenso, col problema che però ci voleva sempre il terzo o il quarto. Allora il terzo o il quarto in tutti i gruppi di consenso sono stati o il presidente o il vicepresidente, o il coordinatore di sezione. Quindi ogni gruppo di consenso è stato formato dai due, minimo, più il presidente o il coordinatore di sezione» (Intervista Colozzi).

[122] La ragione di questa scelta sembra risiedere principalmente nella ricezione delle classificazioni delle riviste da parte della comunità scientifica: «quando sono usciti gli elenchi e sono stati immediatamente delegittimati attraverso due vie: un dibattito che non è arrivato all'insulto, ma quasi, e una serie di ricorsi quasi infinita. Quasi tutti i direttori di rivista hanno fatto ricorso contro la classificazione. A fronte di questo non ce la siamo più sentita di inviare queste nostre graduatorie ai *referee*, per cui abbiamo dato ai *referee* solo i prodotti, e basta. Per cui è peer review, non informed peer review» (Intervista Colozzi).

informazioni incluse nella procedura di *informed* peer review sono discusse di seguito (*cfr*. § 4.3.2.1), per ora è sufficiente sottolineare che questi aspetti avrebbero coinvolto una parte limitata dei prodotti sottoposti a valutazione. Infatti solo 452 articoli sono risultati pubblicati in una delle riviste classificate dal GEV (il 36,6% degli articoli su rivista conferiti e il 10,4% sul totale dei prodotti; *cfr.* Anvur, 2013d, GEV14, p. 29) e soltanto 261 prodotti (il 6% del totale; *cfr.* Anvur, 2013d, GEV 14, Appendice, p. 92) è risultato indicizzato in WoS o Scopus.

All'interno dei rapporti finali non sono disponibili ulteriori informazioni sulle regole di attribuzione utilizzate, tuttavia il documento circa i "Criteri di attribuzione delle classi di merito" (Anvur 2014) riporta anche i criteri di assegnazione utilizzati nel caso di più di due referaggi (tre, quattro e cinque).

I criteri utilizzati per l'assegnazione delle classi di merito sono deducibili dai documenti e dalle interviste[123]:

- nel caso il prodotto ottenga due giudizi concordanti l'assegnazione della classe di merito deve essere approvata dai due esperti valutatori che hanno in carico il prodotto;
- nel caso il prodotto ottenga due giudizi discordanti, di due o tre classi (U), è prevista l'apertura di un gruppo di consenso che può avvalersi di un terzo referaggio;
- nel caso in cui i due giudizi risultino discordanti di una sola classe gli esperti valutatori possono confermare la valutazione corrispondente al punteggio oppure richiedere l'apertura di gruppo di consenso, che ha facoltà di proporre una nuova valutazione, sulla base non solo delle schede compilate dai *referee*, ma anche di ulteriori informazioni disponibili sul prodotto, o di richiedere un terzo referaggio.

E' il caso di sottolineare che perché venisse aperto il gruppo di consenso oppure effettuato il terzo referaggio era sufficiente che anche uno solo degli esperti valutatori che avevano in carico il prodotto obiettasse in merito alla valutazione proposta sulla base dei punteggi. Dalle interviste si evince tuttavia un uso dei gruppi di consenso e delle terze revisioni limitato ai soli casi fortemente discordanti[124].

[123] Sono inoltre stati confermati dalla dottoressa Colizza, assistente GEV di Area 14.

[124] Nelle interviste sono presenti riferimenti a ciascuna delle possibili opzioni: «noi abbiamo fissato delle soglie che permettevano di andare in automatico al numero massimo di prodotti. Abbiamo discusso direttamente quelli che avevano problemi e che quindi non rientravano chiaramente in una soglia o nell'altra soglia. Solo su quelli» (Intervista Colozzi); «quando noi avevamo situazioni molto ben definite, è chiaro non facevamo nessun intervento, insomma se la cosa è evidente non… Veramente quando c'era da decidere se appartenesse a una certa categoria o ad una certa altra categoria, è chiaro che dovevamo ricorrere a un terzo elemento, un terzo valutatore che potesse stabilire un po' il discrimine tra l'uno e l'altra posizione, poi vedevamo anche se era convincente quanto veniva indicato o meno» (Intervista Cipriani); «c'era una proposta, che tra l'altro era una media tra due, insomma non era una proposta così dirimente. Quindi il sistema faceva questa proposta ma in realtà si apriva il gruppo di consenso e quindi i due esperti che avevano in capo il prodotto si gestivano il gruppo di consenso o inviavano a un terzo revisore» (Intervista Blasi); «nel caso le differenze erano di due o tre gradini (eccellente-limitato, oppure eccellente-accettabile), in quel caso abbiamo sempre chiesto un terzo giudizio, da parte di un ulteriore esperto, che abbiamo coinvolto appositamente in questo caso» (Intervista Colozzi). E' inoltre interessante riportare a questo proposito l'opinione del professor Cipriani: «più che una operazione di valutazione è una valutazione procedurale, cioè non siamo noi a valutare, noi valutiamo solo nei casi straordinari in cui c'è discordanza, ma la gran massa delle valutazioni proviene dal mondo accademico e appunto dalla lista dei valutatori che sono stati resi disponibili attraverso le domande o attraverso le nostre indicazioni » (Intervista Cipriani).

Sulla base del documento già citato (Tabella 19, p. 98) è possibile ricostruire un modello per la proposta di assegnazione della classe di merito finale; in Tabella 23 si riporta per semplicità un modello con due revisori.

Tabella 23 - Assegnazione della classe di merito finale in base ai punteggi e alla differenza di classi tra due *referee*

		Punteggio assegnato dal revisore A									
		0	1	2	3	4	5	6	7	8	9
Punteggio assegnato dal revisore B	0	L-L	L-L	L-L	L-L	L-L	L-A	U	U	U	U
	1	L-L	L-L	L-L	L-L	L-L	L-A	U	U	U	U
	2	L-L	L-L	L-L	L-L	L-L	L-A	U	U	U	U
	3	L-L	L-L	L-L	L-L	L-L	L-A	U	U	U	U
	4	L-L	L-L	L-L	L-L	L-L	L-A	U	U	U	U
	5	A-L	A-L	A-L	A-L	A-L	A-A	A-B	A-B	U	U
	6	U	U	U	U	U	B-A	B-B	B-B	B-E	B-E
	7	U	U	U	U	U	B-A	B-B	B-B	B-E	E-E
	8	U	U	U	U	U	U	E-B	E-B	E-E	E-E
	9	U	U	U	U	U	U	E-B	E-E	E-E	E-E

Riportando in tabella le classi di merito anziché i punteggi è possibile ricostruire il modello di assegnazione della classe di merito finale (Tabella 24).

Tabella 24 - Schema di assegnazione delle classi di merito finali in base alle classi conferite dai due revisori

		Classe assegnata dal revisore A			
		L	A	B	E
Classe assegnata dal revisore B	L	L	L	U	U
	A	L	A	A o B	U
	B	U	A o B	B	B
	E	U	U	B	E

La sintesi dei giudizi dei revisori, è il caso di sottolineare, non prevedeva alcun controllo diretto circa l'accordo dei revisori al livello dei singoli criteri[125]. La procedura, infatti, essendo basata sulla somma tra i punteggi totali assegnati ai prodotti e sul grado di accordo sul piano delle classi di merito, non è in grado di identificare le differenze tra i giudizi analitici circa uno stesso prodotto se il loro esito sintetico è lo stesso. Ad esempio se uno dei due revisori dovesse valutare il prodotto come eccellente su rilevanza e internazionalizzazione, ma limitato sull'originalità, e l'altro dovesse ritenerlo eccellente su rilevanza e originalità, ma limitato sull'internazionalizzazione, entrambe le revisioni avrebbero come esito la classe buono, risultando coincidenti.

E' chiaro che un controllo sistematico sul grado di accordo circa i singoli criteri non sarebbe utilizzabile direttamente per la determinazione nell'assegnazione dei prodotti a un terzo revisore, nondimeno potrebbe produrre informazioni estremamente utili per gli EV, tanto riguardo alle valutazioni sintetiche discordanti quanto riguardo alle valutazioni sintetiche concordanti.

Né nel rapporto finale di Area né nelle sue appendici è possibile rinvenire informazioni su quante valutazioni siano state approvate direttamente dagli esperti valutatori, su quante abbiano

[125] Non è chiaro se gli EV, visualizzando la scheda relativa al prodotto e i giudizi dei revisori, potessero accedere ai punteggi assegnati sui singoli criteri. La procedura predefinita per l'assegnazione delle classi si basava esclusivamente sull'accordo relativo all'esito della classificazione, senza segnalare automaticamente eventuali discrepanze nei giudizi analitici, di conseguenza ulteriori controlli erano demandati alla discrezionalità degli EV.

richiesto l'intervento dei gruppi di consenso né di quante abbiano reso necessario un terzo referaggio[126]. Quest'ultimo dato, tuttavia, può essere dedotto a partire da altre due informazioni: il numero di revisioni effettuate e il numero di prodotti sottoposti a peer review.

I prodotti sottoposti a peer review nell'Area 14, includendo dunque anche quelli inviati da altre Aree, sono 4.317 (Tab. 2.3, Anvur, 2013d, GEV 14, p. 21)[127]. Non è tuttavia immediata l'individuazione del numero di revisioni commissionate dal GEV, dato che nel rapporto finale vengono riportate cifre riferibili a revisioni *effettuate*, *inevase* e *rifiutate* (Tabella 25). Considerando il totale delle revisioni (11.949) risulterebbero ben 3.315 revisioni in più rispetto al minimo previsto dalla procedura (8.634, cioè 2 revisioni per ciascuno dei 4.317 prodotti sottomessi a valutazione). Evidentemente, però, dal totale delle revisioni vanno escluse le 2.863 rifiutate per ragioni legate a conflitti di interesse, non competenza, mancanza di tempo o alla lingua; in questo caso il totale delle revisioni sarebbe pari a 9.086, con 452 revisioni supplementari. Una cifra molto più plausibile della precedente. Neppure le revisioni inevase sono state effettuate, nel rapporto infatti si specifica che nel corso della procedura di valutazione si è reso necessario un allargamento del numero di revisori, e che «ciò è dipeso dalle grandi difficoltà indotte dai comportamenti di una minoranza di colleghi stranieri e italiani che, dopo aver dato la loro disponibilità in termini generali, hanno rifiutato la maggior parte (in alcuni casi la totalità) dei prodotti loro proposti o si sono limitati a non esercitare alcuna opzione, lasciando scadere i termini fissati per l'accettazione, o ancora, hanno accettato i prodotti loro inviati, ma non hanno mai compilato la scheda di valutazione […] casi di *malpractice*, che la tabella 2.9. documenta più in dettaglio» (Anvur, 2013d, GEV 14, pp. 26-27).

E' opportuno, dunque, fare riferimento alle sole revisioni effettuate: 8.512, che però sono 122 in meno del minimo previsto dalla procedura (8.634): in altre parole 122 prodotti hanno ricevuto un solo giudizio da parte di un *referee* esterno al GEV[128].

Tabella 25 - Revisioni totali, inevase, effettuate e rifiutate per nazionalità di affiliazione del revisore (Tabella 2.9.1, Anvur, 2013d, GEV14, p. 27)

Revisioni effettuate da revisori con affiliazione:	Revisioni totali	Revisioni effettuate	Revisioni inevase	Revisioni rifiutate per:				Revisioni rifiutate
				Non competenza	Conflitto di interesse	Mancanza di tempo	Lingua	
Italiana	9.138	6.858	257	728	164	1.122	9	2.023
non Italiana	2.811	1.654	317	401	52	316	71	840
Totale	**11.949**	**8.512**	**574**	**1.129**	**216**	**1.438**	**80**	**2.863**

[126] A proposito del numero di terzi referaggi effettuati è interessante riportare direttamente alcuni stralci delle interviste: «ne abbiamo avuti parecchi. Se lei volesse il dato percentuale, se sente la Blasi lei ha tutti questi dati, io non mi ricordo più a distanza ormai di due anni dalla fine. Però forse noi siamo stati, può darsi che sia scritto anche nel rapporto, noi dovremmo essere stati quelli che sono ricorsi di più al terzo giudizio, perché l'asimmetria nelle valutazioni è stata particolarmente significativa» (Intervista Colozzi); «non abbiamo avuto veramente il tempo di analizzare questi dati» (Intervista Blasi).

[127] Ricordiamo infatti che tutti i prodotti valutati in Area 14 sono stati valutati tramite peer review, ma 100 prodotti sottomessi a valutazione da soggetti appartenenti all'Area 14 sono stati inviati per la valutazione ad altri GEV, e che 10 di questi sono stati valutati in bibliometria (Anvur, 2013d, GEV14, p. 19-21).

[128] Altre informazioni presenti nel rapporto risultano coerenti con questo dato. Calcolando infatti il totale della tabella 2.11 che riporta il numero di referaggi per affiliazione (italiana o straniera) del *referee* e per il settore disciplinare si ottiene nuovamente 8.512 (ibidem, p. 28); questi dati inoltre corrispondono con quelli riportati nella tabella 3.2 del rapporto finale dell'Agenzia (Anvur, 2013a, p. 25).

Facendo riferimento a una diversa tabella (Tabella 26) il numero complessivo dei referaggi risulta pari a 8.666, esattamente 32 in più del minimo previsto dalla procedura. La tabella riporta il numero di referaggi gestiti da ciascun Esperto Valutatore (d'ora in avanti EV), intendendo per «"gestiti" che il componente del GEV, seguendo la procedura predisposta dal CINECA, ha accettato di prendersi carico del prodotto (si è "proposto" per il prodotto); dopo l'approvazione ("accettazione") da parte del Presidente, lo ha assegnato a un revisore e, a valutazione avvenuta, ha partecipato al gruppo di consenso che ha convalidato le valutazioni proposte dai *referee*, potendo decidere, in caso di giudizi fortemente divergenti, per la scelta di un terzo revisore» (*ibidem*, p. 9). E' tuttavia plausibile che non siano conteggiati i referaggi effettuati personalmente dagli EV[129].

Tabella 26 - Organizzazione degli esperti in SubGEV, corrispondenti SSD e distribuzione dei prodotti della ricerca *(Adattamento della Tabella 1.2, Anvur, 2013d, GEV14, p. 8)*[130]

Sub-GEV	*EV e SSD di appartenenza*	*SSD di competenza*						*Referaggi gestiti*
SCIENZE POLITICHE	1 – SPS/01	SPS/01	SPS/02	SPS/03	SPS/04	SPS/06		901
	2 – SPS/04	SPS/04	SPS/06	*SPS/09*	*SPS/11*			626
	3 – SPS/01	SPS/01	SPS/02	SPS/03	SPS/04	*SPS/11*		361
	4 - SPS/02	SPS/01	SPS/02	SPS/03				491
	5-SPS/04	***SPS/04***						***210***
	6- SPS/04	SPS/04	SPS/06	*SPS/09*				501
	7-GEV11	***SPS/05***	***SPS/13***	***SPS/14***				***184***
SCIENZE SOCIALI	8 – SPS/07	SPS/07	SPS/09	SPS/11	SPS/12	*SPS/14*		1.411
	9– SPS/07	SPS/07	SPS/08	SPS/09	SPS/10	SPS/11	SPS/12	1.130
	10– SPS/08	SPS/07	SPS/08	SPS/09				765
	11– SPS/09	SPS/09	SPS/10	SPS/12				415
	12– SPS/10	SPS/08	SPS/09	SPS/10				287
	13– SPS/09	SPS/07	SPS/09	SPS/10	SPS/12			511
	14-SPS/08	SPS/07	SPS/08	SPS/09				873
Totale								8.666

Quest'ultima soluzione è la più verosimile. E' inoltre possibile individuare, a latere della tabella, una possibile spiegazione rispetto ai numeri precedentemente riportati; infatti vi si legge che, in mancanza di esperti competenti in alcuni settori[131]: «il GEV ha deciso di chiedere al prof. Andrea

[129] In proposito il Presidente ha sottolineato che i prodotti con valutazioni discordanti di più di una classe sono andati: «tutti al terzo referaggio. Il punto è che il terzo referaggio in alcuni casi è stato uno del GEV, perché quando non siamo riusciti a piazzarlo da nessun'altra parte abbiamo allora sempre fatto la scelta interna della valutazione, piuttosto che non farla. Cioè piuttosto che non farla abbiamo fatto sempre la terza valutazione, in alcuni casi direttamente noi. […]Noi al massimo avremo fatto il 20%, ma neanche. Noi come valutazione interna abbiamo cercato di rimanere sotto il 15%, questa era una soglia che ci eravamo dati. In alcuni casi, se consideriamo solo il sotto-settore disciplinare potrebbe esserci stato uno sforamento, in alcuni casi essere arrivati al 20, in alcuni casi al 10 invece che ha compensato, per cui la media totale su tutti i prodotti è intorno al 15. Se consideriamo il GEV in alcuni casi i rifiuti sono stati di più e allora possiamo essere arrivati al 20% come valutazione interna, ma in alcuni altri siamo stati sotto il 10, perché appunto abbiamo tenuto le valutazioni esterne quasi su tutte» (Intervista Colozzi).

[130] In blu si riportano i settori «subappaltati al GEV 11», in rosso quelli che hanno implicato la gestione di un numero esiguo di prodotti (da 1 a 4 prodotti) da parte dell'Esperto Valutatore (di qui in avanti EV).

[131] I settori in questione sono quelli riportati in blu nella tabella sopra: Storia e Istituzioni delle Americhe (SPS/05), Storia e Istituzioni dell'Africa (SPS/13) e Storia e Istituzioni dell'Asia (SPS/14).

Graziosi, Presidente del GEV di Area 11, la disponibilità a valutare tali prodotti avvalendosi dei *referee* di Area 11» (*ibidem*). E' plausibile che i referaggi corrispondenti a questi SSD, non essendo stati gestiti dal GEV 14, siano stati esclusi dai conteggi precedenti, mentre nella tabella 2.11 (*ibidem*, p. 28) compaiono tutti e tre i settori.

Nel rapporto finale dell'Agenzia viene riportata una tabella sintetica sulle revisioni peer discordanti per 1, 2 e 3 classi per ciascuna delle quattordici Aree (Tabella 27), ma in questa tabella non corrisponde la cifra dei prodotti sottoposti a peer review (4.304 anziché 4317) e la proiezione dei referaggi supplementari non risulta coerente con nessuna delle ipotesi esposte finora. In questo caso infatti i prodotti con referaggi discordanti di più di una classe risultano 970 (833+137), un numero molto lontano sia dall'ipotesi più ampia (470) sia da quella più ristretta (32) in riferimento ai referaggi supplementari.

L'Area delle Scienze Politiche e Sociali presenta una quota abbastanza ridotta di valutazioni concordanti (il 38,5%), e quote decisamente elevate di valutazioni discordanti di una e due classi (rispettivamente il 39,9% e il 19,4%)[132].

Tabella 27 - Numero e percentuali di revisioni peer discordanti per 1, 2 e 3 classi per Area *(Adattamento della Tabella 3.4, Anvur, 2013a, p. 26)*

Area	Prodotti sottomessi 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	
		n	%	n	%	n	%	n	%
1	5.180	3.194	*61,7*	1.314	*25,4*	522	*10,1*	150	*2,9*
2	5.350	2.095	*39,2*	2.125	*39,7*	924	*17,3*	206	*3,9*
3	2.879	1.429	*49,6*	1.054	*36,6*	342	*11,9*	54	*1,9*
4	3.390	2.249	*66,3*	757	*22,3*	318	*9,4*	66	*1,9*
5	4.985	2.328	*46,7*	1.616	*32,4*	855	*17,2*	186	*3,7*
6	10.330	4.839	*46,8*	3.208	*31,1*	1.946	*18,8*	337	*3,3*
7	4.501	2.088	*46,4*	1.519	*33,7*	737	*16,4*	157	*3,5*
8	7.541	3.111	*41,3*	2.710	*35,9*	1.407	*18,7*	313	*4,2*
9	7.351	3.251	*44,2*	2.515	*34,2*	1.274	*17,3*	311	*4,2*
10	13.942	5.556	*39,9*	5.652	*40,5*	2.123	*15,2*	611	*4,4*
11	11.186	4.290	*38,4*	4.470	*40*	2.005	*17,9*	421	*3,8*
12	11.784	4.462	*37,9*	4.765	*40,4*	2.142	*18,2*	415	*3,5*
13	6.277	3.767	*60*	1.642	*26,2*	757	*12,1*	111	*1,8*
14	4.304	1.659	*38,5*	1.675	*38,9*	833	*19,4*	137	*3,2*
Totale	**99.000**	**44.318**	***44,8***	**35.022**	***35,4***	**16.185**	***16,3***	**3.475**	***3,5***

E' evidente che all'interno dei rapporti c'è un problema di coerenza delle informazioni, oltre che di completezza, che si aggiunge ai problemi di trasparenza già rilevati.

[132] Il commento del professor Colozzi a questa questione è molto interessante: «noi dovremmo essere stati quelli che sono ricorsi di più al terzo giudizio, perché l'asimmetria nelle valutazioni è stata particolarmente significativa. Questo l'abbiamo anche sottolineato nel rapporto finale perché ci ha colpito molto, è stato un elemento su cui invitavamo a riflettere, perché evidentemente… evidentemente si fa per dire, probabilmente, un'ipotesi è che non ci sia, almeno non sia definito uno standard. Perché se fosse definito uno standard non potrebbe succedere così spesso che un *referee* valuta eccellente un lavoro che un altro valuta addirittura limitato. Il fatto che questo sia successo molto spesso ci fa capire… ecco ed è successo molto più tra gli italiani che non tra gli stranieri, questo è un altro dato importante, di questo sono sicurissimo» (Intervista Colozzi).

4.3.2.1 La classificazione delle riviste e gli indicatori bibliometrici per l'*informed peer review*

La questione della classificazione in base al merito delle riviste scientifiche di Area è stata una delle più dibattute nel corso della progettazione dell'esercizio di valutazione VQR, nonostante non sia stata in alcun modo utilizzata nel corso dell'esercizio. Nel rapporto finale di Area si legge: «nonostante l'ampio coinvolgimento delle società scientifiche di settore e di un numero consistente di *referee* italiani e stranieri, la pubblicazione della classificazione ha suscitato molte polemiche e alcune proteste da parte di direttori di riviste non soddisfatti del giudizio di non scientificità o di appartenenza ad una delle tre classi stabilite per quelle scientifiche. L'Anvur ha, pertanto, deciso di attivare una apposita procedura per consentire ai direttori delle riviste di richiedere la revisione della classificazione. Ciononostante, la polemica non si è completamente sopita e la classificazione realizzata dai GEV resta un tema controverso» (Anvur, 2013d, GEV 14, p. 29).

La fase di definizione delle classificazioni è senza dubbio quella in cui l'Agenzia ha maggiormente coinvolto le comunità scientifiche di riferimento, interpellando le associazioni e le società scientifiche non solo nella prima fase, ma anche a seguito delle indicazioni ricevute dai *referee* stranieri, in alcuni casi modificando i criteri e proponendo un nuovo elenco. E' stata anche prevista una procedura di revisione che coinvolgesse direttamente i direttori delle riviste[133]. Ciò nonostante non sono mancate le critiche circa criteri e procedure, in particolare, di nuovo, sulla loro trasparenza e controllabilità, vale dunque la pena di trattare brevemente questa questione, e quella più generale dell'informed peer review.

La classificazione delle riviste italiane dell'Area delle Scienze Politiche e Sociali da parte del GEV 14 ha fatto riferimento a una serie di criteri:

- «la regolarità della pubblicazione;
- la composizione del comitato scientifico (presenza di accademici italiani accreditati, presenza di accademici non italiani accreditati);
- la pratica documentata del referaggio anonimo (con presenza di scheda standard, anonimato dell'autore, double peer review);
- la presenza in *Sociological Abstracts, Worldwide Political Science Abstracts, International Political Science Abstracts, Social Services Abstracts, Current Abstracts, SocINDEX;*
- l'indicizzazione presso le principali piattaforme di ricerca bibliografica (*Ebsco Discovery Service, International Bibliography of the Social Sciences* (IBSS), Google Scholar, *ProQuest Summon*, *Casalini Digital Library*, Articoli italiani di periodici accademici (AIDA), Catalogo italiano dei periodici (ACNP), ecc.);
- l'indicizzazione nei database ISI e/o Scopus;
- l'indice H calcolato sulla base del database Google Scholar tramite il software *Publish or Perish*;

[133] A questo proposito, vale la pena di riportare nuovamente le parole del Presidente Colozzi: «quando sono usciti gli elenchi, sono stati immediatamente delegittimati attraverso due vie: un dibattito che non è arrivato all'insulto, ma quasi, e una serie di ricorsi quasi infinita. Quasi tutti i direttori di rivista hanno fatto ricorso contro la classificazione. A fronte di questo non ce la siamo più sentita di inviare queste nostre graduatorie ai *referee*, per cui abbiamo dato ai *referee* solo i prodotti, e basta. Per cui è peer review, non informed peer review» (Intervista Colozzi).

- la presenza nelle biblioteche universitarie e dei centri di ricerca specializzati» (Anvur, 2013d, GEV14, Appendice, pp. 84-85).

In aggiunta a questi criteri il GEV 14 ha deciso:

- «di non assegnare una classificazione alle riviste sociologiche nate dopo il 2009;
- di escludere dalla prima fascia le riviste che siano espressione di un'unica sede universitaria e non contemplino una significativa presenza di contributi "esterni";
- di inserire negli elenchi riviste che compaiono anche negli elenchi di altre aree scientifiche» (Anvur, 2013d, GEV14, Appendice, p. 85).

Tralasciando per un attimo queste ultime specificazioni e mettendo a fuoco i criteri, è possibile tentare di individuare alcune linee essenziali nella classificazione delle riviste. Alcuni di questi criteri sembrano riferibili alla solidità della rivista, in particolare il primo, altri anche alla sua autorevolezza, come la composizione del comitato scientifico e la pratica documentata del referaggio anonimo, tuttavia la maggior parte dei criteri è connessa alla visibilità e alla diffusione (nazionale e internazionale) della rivista. Infatti la presenza nelle biblioteche universitarie e dei centri di ricerca specializzati, nei principali cataloghi di abstracts, nelle piattaforme di ricerca bibliografica, nei database citazionali WoS e Scopus, sono tutti indicatori di visibilità e accessibilità. Infine uno dei criteri mira a tenere conto dell'impatto della rivista: l'indice H calcolato sulla base dei dati forniti da Google Scholar utilizzando il software *Publish or Perish* (sul punto si tornerà fra poche righe).

Non conoscendo la definizione operativa di nessuno di questi criteri, né l'algoritmo che ha condotto alla classificazione delle riviste, non è possibile esaminare le scelte procedurali, ma è comunque opportuno avanzare alcune osservazioni di carattere generale.

Indubbiamente la solidità della rivista è un elemento da tenere in considerazione nella formulazione di un ranking, tuttavia ci sembra che assumere la regolarità della sua pubblicazione come unico criterio possa lasciar fuori una serie di elementi altrettanto interessanti, come ad esempio il trend degli abbonamenti, oltre che presentare dei problemi di affidabilità[134].

In riferimento alla composizione del comitato scientifico sembra riduttivo limitare la questione ai membri accademici accreditati italiani e stranieri, anche se non è semplice individuare indicatori alternativi della qualità del comitato scientifico.

Indubbiamente uno dei criteri più rilevanti dovrebbe essere la pratica del referaggio anonimo, soprattutto se documentata, infatti «è compito dei *referee* nel processo di revisione, come *gatekeepers* della scienza, di raccomandare per la selezione della migliore ricerca scientifica sotto la condizione di risorse scarse (come lo spazio limitato nelle riviste, fondi limitati) (Hackett e Chubin, 2003). Inoltre, i revisori sono tenuti a scoprire errori di pubblicazioni scientifiche e riconoscere cattiva condotta scientifica (Smith 2006)[135]» (Bronmann, 2011, p. 24). La presenza di una pratica di referaggio anonimo dovrebbe dunque rappresentare un indicatore ragionevole di qualità in riferimento alla produzione scientifica pubblicata nella rivista, e alla solidità della rivista stessa.

La visibilità della rivista, infine, può essere assunta come un elemento di qualità e la presenza nelle biblioteche universitarie e nei centri di ricerca specializzati, nei principali cataloghi di abstracts, nelle piattaforme di ricerca bibliografica, nei database citazionali WoS e Scopus, tuttavia a questa

[134] E' stato sottolineato come al di là del numero e dell'annata riportate in copertina, a volte le riviste scientifiche escano con mesi di ritardo (http://www.roars.it/online/classificazione-riviste-esigiamo-trasparenza-lettera-aperta-al-presidente-dellAnvur-e-al-presidente-del-gev-14/).

[135] Traduzione dall'originale in lingua inglese.

dimensione andrebbe attribuito un peso adeguato. E' rilevabile che quattro criteri su otto siano legati a questo aspetto, ma non è possibile conoscere quale sia il peso assegnato a ciascuno di essi nella determinazione della classificazione della rivista.

Il criterio meno convincente e più problematico è: «l'indice H calcolato sulla base del database Google Scholar tramite il software *Public or Perish»* (Anvur, 2013d, GEV14, Appendice, p. 85). L'impatto della rivista fornisce una quota fondamentale dell'informazione relativa alla sua qualità, ma gli indici bibliometrici più comunemente riferiti a questa dimensione sono in effetti l'*impact factor* di Web of Science e l'indice SJR di Scopus. L'H-index è stato proposto nel 2005 dal fisico J.H. Hirsch (di qui la denominazione di Hirsch Index o H-index), con lo scopo di quantificare la prolificità e l'impatto del lavoro di singoli ricercatori, tenendo conto tanto del numero delle loro pubblicazioni quanto del numero di citazioni da esse ricevute (Hirsch, 2005). L'indice non è affatto complesso: un ricercatore ha un valore dell'H-index pari a *n* se ha pubblicato almeno *n* lavori, ciascuno dei quali è stato citato almeno *n* volte, eppure ha scatenato un ampio dibattito. Trascureremo la rendicontazione del dibattito che ha fatto seguito alla sua comparsa (per una rassegna: Bornmann e Daniel, 2009), ma è opportuno sottolineare che l'indice H è stato pensato per i singoli ricercatori, e solo adattato al fine di misurare l'impatto delle riviste (Braun, Glänzel e Schubert, 2006). Andrebbe valutata l'opportunità di utilizzare proprio quest'indice, molto discusso e poco consolidato, con riferimento alle riviste.

Tutti gli indici bibliometrici risentono fortemente della base di dati su cui sono calcolati, in ragione della loro differente copertura, cioè del numero e della varietà delle riviste indicizzate (*cfr.* Capitolo 5). Il GEV 14 fa riferimento a Google Scholar[136]. Il punto di forza di Scholar è senza dubbio la copertura: mentre infatti WoS e Scopus sono database bibliometrici con una copertura selettiva delle riviste[137] Scholar presenta una copertura pseudo-universale. Il suo punto di debolezza tuttavia è l'affidabilità: non è in grado di identificare univocamente né i prodotti né i loro autori, né di controllare il tipo di prodotto includendo nei conteggi i testi più vari[138]. Inoltre, i dati di Scholar provengono dal Web, dunque la sua copertura risulta limitata ai prodotti presenti on-line al momento dell'interrogazione, fortemente instabile oltre che condizionata dal periodo della pubblicazione[139], di conseguenza gli indicatori calcolati sulla base di Google Scholar risultano soggetti a variazioni anche nel breve termine (Jacsó, 2005, Bakkalbasi *et al.* 2006, Bar-Ilan, 2008). Infine non di rado in Scholar informazioni come anno di pubblicazione, denominazione della rivista ed editore risultano mancanti, incoerenti in diversi record o errati. E' stato sottolineato, proprio in riferimento all'H-index, che la scelta del database citazionale da utilizzare non è affatto indifferente rispetto all'esito del calcolo, e gli studi condotti sul confronto tra i database: «mostrano un'elevata

[136] Scholar è la versione di Google per la ricerca scientifica, e copre diversi prodotti della ricerca, principalmente ma non esclusivamente articoli, ma anche relazioni, libri, tesi e *working papers*, presenti sui siti di case editrici, università, istituti di ricerca e associazioni sia accademiche che professionali. Scholar è utilizzabile per l'analisi citazionale perché è in grado di rilevare il numero di volte in cui un dato documento viene citato in altri documenti.

[137] Le riviste, tanto per WoS quanto per Scopus, sono selezionate in base a una procedura di valutazione da parte di esperti, basata sia su misure quantitative che su misure qualitative (*cfr*. Capitolo 5).

[138] A titolo di esempio si pensi agli abstract, agli interventi sui siti delle associazioni, alle slides e agli appunti delle lezioni: «la composizione esatta di questo materiale citazionale dovrebbe essere esaminata più attentamente in modo che gli studiosi abbiano una chiara idea di ciò che è o non è incluso nelle estrazioni di Google Scholar» (Bakkalbasi et al., 2006, p. 7/8, Traduzione dall'originale in lingua inglese).

[139] Risulta del tutto inaffidabile per prodotti pubblicati prima del 1990 (Below, 2005; Meho e Yang, 2007).

somiglianza tra Scopus e WOS e somiglianze più piccole tra Google Scholar e gli altri strumenti, il che indica che la copertura di Google Scholar è notevolmente diversa da quella di WOS e Scopus[140]» (Bar-Ilan, 2008, p. 261). Nonostante non manchino opinioni possibiliste su un futuro utilizzo di Scholar come database citazionale (Noruzi, 2005), la posizione più diffusa è sostanzialmente critica: «sfortunatamente Google Scholar procura una cattiva fama all'analisi citazionale autonoma. Mostra una mancanza di competenza e comprensione delle problematiche basilari nell'analisi citazionale[141]» (Jacsó, 2005, p. 1546).

Attraverso l'applicazione Google Scholar Citation[142], Scholar può essere usato direttamente per il calcolo di alcuni indicatori citazionali (autenticandosi come autore del documento), ma per estrarre dati relativi a prodotti di cui non è possibile riconoscersi come autori è necessario utilizzare interfacce esterne. Ne esistono diverse e quella indicata dal GEV 14, *Publish or Perish*, è solo una di queste. *Publish or Perish* consente di produrre una serie di indicatori bibliometrici, tra cui l'H-index, e, in una certa misura, di controllare e pulire i dati di partenza estratti da Scholar, tuttavia non elimina tutti i problemi connessi alla scelta di questa base di dati. Pur permettendo un maggiore controllo sull'identificazione dell'autore e del prodotto, il software non è in grado di distinguere i prodotti dalle citazioni, oppure di integrare o correggere i dati disponibili, dunque non risolve i problemi legati all'affidabilità del conteggio delle citazioni: «a questo proposito, *Publish or Perish* è alla mercé di Google Scholar[143]» (Jacsó, 2009, p. 1198).

La scelta di inserire tra i criteri di classificazione delle riviste proprio l'H-index calcolato da *Publish or Perish* su Scholar risulta in sintesi quantomeno discutibile tanto dal punto di vista dell'adeguatezza dell'indicatore quanto dal punto di vista dell'adeguatezza della base di dati.

Veniamo alle ulteriori decisioni prese dal GEV 14 in riferimento alla classificazione delle riviste (Anvur, 2013d, GEV14, Appendice). La prima decisione segnalata è quella di «non assegnare una classificazione alle riviste sociologiche nate dopo il 2009» (*ibidem*, p. 85), inoltre tutte le riviste nate dopo il 2008 vengono segnalate, e la loro classificazione è provvisoria. Dato che il periodo di riferimento dell'esercizio di valutazione è il settennio 2004/2010 è comprensibile che la procedura non includa le riviste nate in un periodo troppo a ridosso della scadenza: infatti per queste riviste risulta difficile valutare non solo l'impatto, ma anche la visibilità e la solidità.

Una seconda decisione del GEV prevede di «escludere dalla prima fascia le riviste che siano espressione di un'unica sede universitaria e non contemplino una significativa presenza di contributi "esterni"» (*ibidem*). Questa scelta non viene argomentata, tuttavia sembra sottintendere che riviste che siano espressione di un'unica abbiano una qualità inferiore rispetto a riviste con le stesse caratteristiche ma con contributi esterni. E' in effetti improbabile che una rivista "chiusa" riesca a ottenere un impatto e una visibilità elevati, d'altro canto se riesce ad ottenerli l'unica spiegazione è nella qualità dei suoi contenuti, inoltre una rivista con una procedura di peer review anonima difficilmente può essere tanto chiusa da risultare «espressione di un'unica sede universitaria». Insomma o i criteri alla base della classificazione non sono ritenuti abbastanza stringenti da escludere dalla prima fascia le riviste di respiro corto, oppure la scelta è stata presa per evitare di favorire alcune sedi piuttosto che altre.

[140] Traduzione dall'originale in lingua inglese.
[141] Traduzione dall'originale in lingua inglese.
[142] Si veda in proposito la pagina: https://scholar.google.it/intl/it/scholar/citations.html.
[143] Traduzione dall'originale in lingua inglese.

L'ultima decisione segnalata dal GEV è di «inserire negli elenchi riviste che compaiono anche negli elenchi di altre aree scientifiche[144]» (*ibidem*), in particolare nei casi in cui i contributi dei ricercatori di Area 14 risultino significativi, tanto in termini di prodotti pubblicati quanto in termini di presenza all'interno degli organismi direttivi (comitati editoriali, ecc.). Scelta che appare del tutto ragionevole e condivisibile, specialmente tenendo conto della natura ampiamente interdisciplinare della ricerca nelle Scienze Politiche e Sociali.

Si è già sottolineato che solo 452 articoli sono risultati pubblicati in una delle riviste classificate dal GEV (il 36,6% degli articoli su rivista conferiti, cioè il 10,4% sul totale dei prodotti; *cfr.* Anvur, 2013d, GEV14, p. 29). Osservando la Tabella 28 non sembra che la classe di merito assegnata agli articoli pubblicati su riviste non classificate abbia risentito della mancanza di questa informazione. Le percentuali di "buono" e "accettabile" sono molto simili nei due gruppi di articoli, una differenza leggermente più marcata si ha in riferimento ai "limitato", e soprattutto agli "eccellente". Il GEV 14 ha spiegato la percentuale più elevata di prodotti eccellenti tra gli altri articoli conferiti facendo riferimento alla presenza in questo gruppo degli articoli pubblicati sulle riviste internazionali, che non rientrano nella classificazione del GEV.

Tabella 28 -Articoli su rivista censita nel rating del GEV: confronto tra classificazione riviste e valutazione peer (adattamento della tabella 2.13; Anvur, 2013d, GEV 14, p. 30)

		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
Valori assoluti	Articoli scritti su riviste classificate	23	131	149	149	**452**
	Altri articoli conferiti	125	250	214	195	**784**
	Totale articoli su rivista	**148**	**381**	**363**	**344**	**1.236**
		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
% di riga	Articoli scritti su riviste classificate	15,5	34,4	41,1	43,3	**36,6**
	Altri articoli conferiti	84,5	65,6	59,0	56,7	**63,4**
	Totale articoli su rivista	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
% di colonna	Articoli scritti su riviste classificate	5,1	29,0	33,0	33,0	**100,0**
	Altri articoli conferiti	15,9	31,9	27,3	24,9	**100,0**
	Totale articoli su rivista	**12,0**	**30,8**	**29,4**	**27,8**	**100,0**

Il report di Area riporta anche il confronto tra la classe di merito finale assegnata agli articoli e la fascia della rivista in cui sono stati pubblicati (Tabella 29). La tabella mostra un tendenziale, anche se scarso, accordo tra le due classificazioni, e il GEV ne conclude che: «generalizzando si può dire che la probabilità che un articolo pubblicato su una rivista di classe A sia valutato eccellente è leggermente più alta di quella di un articolo pubblicato su riviste classificate in fascia B o C o che la concordanza tra le valutazioni *peer* e la nostra classificazione è risultata piuttosto bassa, ma non trascurabile» (Anvur, 2013d, GEV14, p. 32).

[144] In questo caso negli elenchi le riviste venivano accostate alla sigla (*int*), indicante riviste intersettoriali o interarea.

Tabella 29 -Articoli su rivista censita nel rating del GEV: confronto tra classificazione riviste e valutazione peer (adattamento della tabella 2.14; Anvur, 2013d, GEV 14, p. 32)

		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
Valori Assoluti	**A**	17	68	89	76	**250**
	B	5	59	47	59	**170**
	C	1	4	13	14	**32**
	Totale	**23**	**131**	**149**	**149**	**452**
		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
% di riga	**A**	6,8	27,2	35,6	30,4	**100,0**
	B	2,9	34,7	27,6	34,7	**100,0**
	C	3,1	12,5	40,6	43,8	**100,0**
	Totale	**5,1**	**29,0**	**33,0**	**33,0**	**100,0**
		Classe peer convalidata				Totale
		E	**B**	**A**	**L**	
% di colonna	**A**	73,9	51,9	59,7	51,0	**55,3**
	B	21,7	45,0	31,5	39,6	**37,6**
	C	4,3	3,1	8,7	9,4	**7,1**
	Totale	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

Nonostante la procedura non prevedesse che i revisori venissero informati della classe della rivista in cui i prodotti erano stati pubblicati, non è chiaro se e quanto l'informazione relativa alla classe della rivista precedentemente pubblicata dal GEV possa avere influito sull'assegnazione finale della classe di merito. Verosimilmente questo confronto dovrebbe fornire indicazioni sulla validità delle due classificazioni e l'ottica sottintesa dovrebbe essere quella della validità convergente, il fatto però che la classe di merito finale possa essere stata assegnata anche sulla base delle indicazioni fornite dalla classificazione delle riviste crea una sorta di cortocircuito che rende difficilmente interpretabili i risultati del confronto.

La principale indicazione che se ne può trarre, e che viene messa in evidenza dallo stesso GEV, è che «la possibilità/opportunità di sostituire la valutazione *peer* degli articoli con quella basata sulla classe della rivista su cui sono pubblicati sia difficile da sostenere» (*ibidem*).

La procedura di valutazione dei prodotti, anche nelle Aree non bibliometriche, era affiancata dal confronto sperimentale delle valutazioni peer con l'esito delle valutazioni bibliometriche, nel caso delle Scienze Politiche e Sociali il confronto è stato operato per gli articoli indicizzati nei database *Social Sciences Citation Index e Arts and Humanities Citation Index* di Wos o *Social Sciences & Humanities* di Scopus (Anvur, 2013d, GEV 14, Appendice, p. 74). In Area 14 soltanto 261 prodotti (il 6% del totale; *ibidem* p. 92) è risultato indicizzato in almeno una delle due banche dati, dunque solo per questa quota di prodotti erano disponibili dati bibliometrici.

Lo scarso numero di prodotti indicizzati non stupisce, non solo per ragioni legate alle caratteristiche della produzione scientifica delle discipline di Area 14 (discussa di seguito), ma anche per le caratteristiche dei database bibliometrici utilizzati. Notoriamente i due database citati favoriscono prodotti in lingua inglese (van Leeuwen *et al.* 2001; Archambault *et al.* 2006). La stima in WoS per la produzione scientifica in lingua italiana è di una sottorappresentazione dell'80% per le scienze sociali, dell'83% per le scienze naturali (Archambault *et al.* 2006).

Dei 261 prodotti indicizzati 149 (il 57%) risultano indicizzati in WoS, 221 (l'85%) in Scopus, e 126 in entrambe le banche dati (*ibidem*, p. 108): «confrontando ISI WoS e Scopus, quest'ultima banca dati offre una maggiore copertura e valutazioni mediamente più generose» (*ibidem*, p. 111).

Copertura dei database ed esito delle valutazioni non sono indipendenti: maggiore è la copertura del database, più alto potrà essere il numero di citazioni individuate in riferimento a una data rivista o a un dato articolo, più alto è il numero di elementi da includere nelle distribuzioni cumulate, più asimmetrica può rivelarsi la distribuzione (*cfr.* Capitolo 5). Non mancano contributi circa la più ampia copertura offerta dal database di Scopus per le Scienze Sociali (Norris e Oppenhaim, 2007), ma il continuo allargamento e aggiornamento dei database rende difficile trarre delle indicazioni conclusive dalla letteratura, inoltre per l'Area 14 la procedura di valutazione prevedeva che ai prodotti presenti su entrambi i database fosse assegnata la classe derivata da WoS (*ibidem*, p. 103).

Il GEV segnala che le informazioni bibliometriche sono derivate «da un algoritmo che tiene conto, in misura diversa a seconda della data di pubblicazione dell'articolo, sia del numero di citazioni che dell'indicatore bibliometrico della rivista ospitante» (*ibidem*). Così come per la maggior parte delle aree bibliometriche, per l'Area delle Scienze Politiche e Sociali l'algoritmo di classificazione prevede il confronto tra un indicatore citazionale e un indicatore relativo all'impatto della rivista in cui l'articolo è stato pubblicato (ciascuno ricondotto a una delle quattro classi di merito in base al confronto con la distribuzione cumulata delle pubblicazioni nella stessa *subject catgory* o classe dell'ASJC) e l'attribuzione della classe di merito finale sulla base di due matrici quadrate (Tabella 30).

Tabella 30 - Matrici di corrispondenza tra classi iniziali di IF e citazioni e classe finale VQR (adattamento delle tabelle A.13 e A.14; Anvur, 2013d, GEV 14, Appendice, p.102)

2004-2008					
		Impatto			
		E	B	A	L
Citazioni	E	E	E	B	B
	B	E	B	A	A
	A	B	A	A	L
	L	L	L	L	L

2009-2010					
		Impatto			
		E	B	A	L
Citazioni	E	E	B	B	L
	B	E	B	A	L
	A	E	B	A	L
	L	B	A	A	L

A proposito dell'assegnazione delle classi, il GEV segnala che: «trattandosi di prodotti e riviste con indicatori citazionali bassissimi, le soglie bibliometriche che decretano l'assegnazione di un prodotto alle classi sono numeri molto piccoli» (*ibidem*, p. 104, nota 7).

Il confronto condotto, a scopo esplorativo, tra l'esito della classificazione ottenuta tramite analisi bibliometrica e quello della procedura di peer review viene riportato dal GEV 14 nell'Appendice C al rapporto finale di Area[145] (Tabella 31).

La valutazione bibliometrica è allo stesso tempo più generosa e più severa della valutazione peer: tanto i prodotti classificati come eccellenti quanto quelli classificati come limitati risultano infatti molto più numerosi, a scapito soprattutto delle classi centrali. Dopo quanto osservato circa la scheda di valutazione utilizzata e i possibili effetti di distorsione cui avrebbe potuto dar luogo non stupisce che la procedura di revisione dei pari sovradimensioni le classi centrali. Complessivamente il 30,3% delle valutazioni risulta perfettamente concordante, ma la valutazione peer risulta più generosa di quella bibliometrica nel 42,2% (106 di 251) dei casi e più severa solo nel 27,5% (69 di 251). Come evidenziato dal GEV: «quando la valutazione è discordante, la bibliometria è più severa e declassa 6 volte su 10 il giudizio dei pari. Questo è più evidente soprattutto nei salti di due classi e

[145] Il GEV sottolinea che da questa analisi sono stati esclusi i 10 prodotti "extra-GEV" che, essendo stati rimessi alla valutazione di GEV esclusivamente bibliometrici, non hanno ricevuto nessuna valutazione peer.

sono solo i 3 casi su 261 prodotti in cui la bibliometria fa passare un prodotto dalla classe di valutazione "Limitato" alla classe "Eccellente"» (Anvur, 2013d, GEV 14, Appendice, p. 104).

In questo caso, come nel caso della classificazione delle riviste, non è possibile conoscere se e quanto i revisori abbiano utilizzato le informazioni bibliometriche disponibili al fine di valutare i prodotti della ricerca. Vale la pena sottolineare però che mentre la classificazione delle riviste messa a punto dal GEV era facilmente reperibile, le classi di merito assegnate ai prodotti dalla procedura di valutazione tramite analisi bibliometrica non erano note ai revisori né potevano essere ricostruite agevolmente.

Tabella 31 - Confronto tra le valutazioni bibliometriche e peer (adattamento della tabella A.8146; Anvur, 2013d, GEV 14, Appendice, p.104)

			Valutazione peer				
Valori assoluti			E	B	A	L	Totale
	Valutazione bibliometrica	E	18	35	11	3	67
		B	8	33	12	4	57
		A	3	22	7	4	36
		L	7	47	19	18	91
		Totale	36	137	49	29	251
% di riga			Valutazione peer				
			E	B	A	L	Totale
	Valutazione bibliometrica	E	26,9	52,2	16,4	4,5	100,0
		B	14,0	57,9	21,1	7,0	100,0
		A	8,3	61,1	19,4	11,1	100,0
		L	7,7	51,6	20,9	19,8	100,0
		Totale	14,3	54,6	19,5	11,6	100,0
% di colonna			Valutazione peer				
			E	B	A	L	Totale
	Valutazione bibliometrica	E	50,0	25,5	22,4	10,3	26,7
		B	22,2	24,1	24,5	13,8	22,7
		A	8,3	16,1	14,3	13,8	14,3
		L	19,4	34,3	38,8	62,1	36,3
		Totale	100,0	100,0	100,0	100,0	100,0

Le ragioni per cui l'applicazione della valutazione bibliometrica alla produzione scientifica delle Scienze Sociali è molto discussa non sono legate esclusivamente ai limiti dei database disponibili ma anche e soprattutto alle specifiche caratteristiche di queste discipline. Hicks (1999) individua tre ragioni principali: la frammentarietà della produzione scientifica, le caratteristiche delle abitudini citazionali e la rilevanza delle letterature nazionali.

La letteratura nelle scienze sociali è estremamente frammentata, non solo perché i significati dei termini variano nel tempo e nello spazio, ma soprattutto perché la maggior parte di queste discipline sono multi-paradigmatiche, includono cioè differenti approcci teorici e metodologici che non di rado risultano in competizione tra loro (Hicks, 1999). Questa situazione si riflette nell'assenza di un solido nucleo di riviste di riferimento, dunque nella obbiettiva difficoltà di selezionarne alcune come particolarmente rappresentative (Archambault *et al.* 2006).

Inoltre le pratiche comunicative nelle scienze sociali sono molto diverse da quelle vigenti nel campo delle scienze dure, la varietà delle forme di rendicontazione dei risultati di ricerca è molto più

[146] Nella tabella originale vengono riportate esclusivamente le percentuali di riga.

elevata e in quest'ambito le citazioni di articoli su rivista sono meno del 25% (Lariviere *et al.* 2006)[147]. Per questa ragione «nelle scienze sociali e umane, l'applicazione dei tassi di citazione delle riviste come un surrogato per il totale di tassi di citazione è più probabile che risulti fuorviante rispetto a quanto lo sia nelle altre scienze» (Bourke *et al.* 1996, p. 54[148]).

Nelle scienze dure i problemi di ricerca sono universali, mentre nelle scienze sociali risultano spesso strettamente connesse a contesti locali, e non di rado i concetti utilizzati e le questioni affrontate possono essere pienamente compresi solo nell'ambito culturale in cui hanno avuto origine. Per questa ragione il pubblico di riferimento delle pubblicazioni è spesso limitato al livello nazionale o regionale (Glänzel*,* 1996; Hicks, 1999 e 2004) e i ricercatori di queste discipline sono più disposti a pubblicare i propri lavori nella lingua nazionale e anche su riviste con una diffusione limitata. Come evidenziato da van Raan «la valutazione bibliometrica della performance di ricerca è basata su un assunto importate: il lavoro da valutare deve essere pubblicato nella letteratura delle riviste aperta e internazionale. Ciò vuol dire che gli indicatori bibliometrici sono applicabili nelle scienze naturali e della vita. Nelle scienze applicate e ingegneristiche come nelle scienze sociali e del comportamento o nelle scienze umanistiche le riviste internazionali spesso non sono il canale principale di comunicazione, dunque la valutazione bibliometrica diventa problematica[149]» (van Raan, 1996, p. 404)

In conclusione se è vero che l'unico uso ragionevole degli indicatori bibliometrici è quello informativo nell'ambito dell'*informed* peer review è anche vero che sarebbe auspicabile utilizzare database appropriati (van Raan, 1996) e che per la produzione scientifica italiana in generale, e in particolare quella riferibile all'area delle scienze sociali, un database del genere non è disponibile. La scelta del GEV di Area 14 di utilizzare una procedura di peer review non informata è dunque indubbiamente la scelta più corretta.

E' vero che nel documento sui criteri della valutazione nell'Area 14 si segnala che nell'assegnare la classe di merito si sarebbe tenuto conto di tutte le informazioni disponibili. Riguardo le monografie, ad esempio, vengono elencate una serie di caratteristiche della collana editoriale in cui il testo è stato pubblicato: «si valuterà l'esistenza di un comitato editoriale, di procedure trasparenti di revisione per decidere sulla pubblicazione, della diffusione a livello nazionale e internazionale dei prodotti dell'editore, di recensioni dell'opera pubblicate su riviste internazionali, e ogni altro elemento atto a fornire indicazioni utili sulla qualità e impatto dell'opera» (Anvur, 2013d, GEV 14, Appendice, p. 74). Tuttavia il passo riportato è l'unico in cui si fa riferimento a

[147] Una meta-analisi condotta da Hicks alla fine degli anni '90 indicava che «i libri comprendono almeno il 40% e forse fino al 60% della letteratura scientifico-sociale. I libri sono molto citati individualmente e, collettivamente, rappresentano circa il 40% delle citazioni» e che «le citazioni da e verso i libri sono distribuite in modo diverso dalle citazioni da e verso gli articoli su rivista» (Hicks, 1999, p. 201; tradotto dall'originale in lingua inglese). E' plausibile che oggi questi dati siano datati e che i rapporti tra i vari elementi si siano modificati, è tuttavia evidente che la centralità della produzione scientifica al di fuori delle riviste abbia ancora un ruolo centrale nelle scienze umane e sociali.

[148] Traduzione dall'originale in lingua inglese.

[149] Traduzione dall'originale in lingua inglese.

queste informazioni, né nel report finale né nelle altre appendici sono disponibili ulteriori dettagli o notizie in proposito[150].

4.3.3 La sintesi e le classi di merito nell'Area delle Scienze Chimiche

L'assegnazione della classe di merito finale nell'Area 3 non prevedeva necessariamente la discussione all'interno dei gruppi di consenso. Ciascun articolo era assegnato a un solo membro del GEV sulla base del Settore Scientifico Disciplinare, e l'EV aveva il solo compito di inviare ai revisori i prodotti estratti come appartenenti alla quota 10% (tutte le Aree hanno sottoposto a peer review il 10% dei prodotti) oppure classificati come *undecided* (Anvur, 2013d, GEV3, p. 11). La classe di merito era dunque assegnata direttamente sulla base delle matrici di corrispondenza riportate nel documento sui criteri di valutazione (Anvur, 2013d, GEV3, Appendice B). Queste matrici stabiliscono una corrispondenza tra la classe assegnata in ragione della posizione sulla distribuzione cumulata di ciascuno degli indicatori utilizzati e la classe di merito finale, prevedendo la valutazione peer in caso di forti discordanze (IR, *informed review*).

L'attribuzione della classe finale non pone problemi nel caso in cui le due classificazioni risultino perfettamente concordanti (Tabella 32).

Tabella 32 – Assegnazione delle classi di merito finali per classi iniziali concordanti su *Impact factor* e citazioni (adattamento da Anvur, 2013d, GEV3, Appendice, p. 22)

		Impact factor			
		L	A	B	E
Citazioni	L	L			
	A		A		
	B			B	
	E				E

La classificazione dei casi al di fuori della diagonale principale avveniva sulla base di due matrici di corrispondenza, differenziate per anno di pubblicazione del prodotto. Infatti «per i lavori più recenti è stato dato un peso maggiore al ranking della rivista piuttosto che al numero di citazioni, giacché la storia citazionale di questi prodotti non è ancora consolidata. Questa scelta ha comportato l'uso di due diverse matrici di corrispondenza (IF-numero di citazioni) una per il periodo 2004-08 e l'altra per il biennio 2009-2010» (Anvur, 2013d, GEV3, p. 12; *cfr.* Tabella 33).

[150] La risposta del professor Colozzi a una domanda sulla eventuale disponibilità di informazioni sulle caratteristiche della casa editrice è chiaramente negativa: «no, anche perché contestualmente, cioè durante la VQR ancora non conclusa, l'Anvur ha nominato una commissione libri e riviste per ciascuna Area, che era incaricata esattamente di questi scopi. Cioè creare un elenco validato, anzi una classifica validata delle riviste nazionali ed internazionali di Area e cominciare ad indicare i criteri di classificazione delle case editrici. Questa commissione è già scaduta, siamo alla seconda commissione, ma ancora i criteri non si son visti» (Intervista Colozzi).

Tabella 33 - Assegnazione delle classi di merito finali in base alla classi iniziali su *Impact factor* e citazioni, per anno di pubblicazione del prodotto (adattamento da Anvur, 2013d, GEV3, Appendice, pp. 23-24)

2004-2008					
		Impact factor			
		L	A	B	E
Citazioni	L	L	L	L	IR
	A	A	A	A	IR
	B	IR	B	B	B
	E	IR	E	E	E

2009-2010					
		Impact factor			
		L	A	B	E
Citazioni	L	L	IR	IR	IR
	A	L	A	B	E
	B	L	A	B	E
	E	IR	IR	IR	E

L'*impact factor* della rivista assume un peso estremamente limitato nella determinazione della classe di merito per gli articoli pubblicati tra il 2004 e il 2008, determinando non una modifica della classe di merito ma l'invio a peer review (IR) per le pubblicazioni:

a- classificate come *eccellenti* dall'IF ma *limitate* dall'indicatore citazionale;
b- classificate come *eccellenti* dall'IF ma *accettabili* dall'indicatore citazionale;
c- classificate come *limitate* dall'IF ma *eccellenti* dall'indicatore citazionale;
d- classificate come *limitate* dall'IF ma *buone* dall'indicatore citazionale (Tabella 33).

In riferimento agli articoli pubblicati tra il 2009 e il 2010 è l'indicatore citazionale ad avere un peso ridotto. L'*impact factor* della rivista ospitante determina direttamente la classificazione salvo alcuni casi per cui è prevista la valutazione peer, quelli in cui i prodotti sono:

a- classificati come *eccellenti* dall'IF ma *limitate* dall'indicatore citazionale;
b- classificati come *buoni* dall'IF ma *limitate* dall'indicatore citazionale;
c- classificati come *accettabili* dall'IF ma *limitate* dall'indicatore citazionale;
d- classificati come *buoni* dall'IF ma *eccellenti* dall'indicatore citazionale;
e- classificati come *accettabili* dall'IF ma *eccellenti* dall'indicatore citazionale;
f- classificati come *limitati* dall'IF ma *eccellenti* dall'indicatore citazionale (Tabella 33).

La preferibilità dell'*impact factor* all'indicatore citazionale per le pubblicazioni più recenti è una delle poche scelte argomentate (seppur brevemente) nella rendicontazione della VQR. L'affidabilità di un indicatore citazionale è in effetti estremamente legata alla "maturità" delle informazioni: «nel caso del numero delle citazioni, si può dire che questo diventi relativamente meno affidabile nella rappresentazione della qualità di un articolo con la diminuzione del tempo trascorso tra la data di pubblicazione e la data di osservazione del numero di citazioni ricevute[151]» (Abramo *et al.* 2010, p. 832). Una questione da non sottovalutare se si considera che la valutazione della ricerca, soprattutto se mirata all'allocazione di risorse, dovrebbe basarsi sui risultati più recenti ottenuti dalle strutture. Abramo e il suo gruppo su questo aspetto concludono che «l'*impact factor* può così essere un predittore del reale impatto di un articolo, e forse migliore delle citazioni[152]» (*ibidem*).

Nei termini di Lazarsfeld (1958) l'*impact factor* e il SJR si configurano come indicatori predittivi, mentre il numero delle citazioni può essere considerato come un indicatore espressivo[153]. Questa differenza si riflette in parte nell'attribuzione delle classi di merito nelle due matrici, dato che

[151] Traduzione dall'originale in lingua inglese.
[152] Traduzione dall'originale in lingua inglese.
[153] Indicatori espressivi e indicatori predittivi non vengono definiti da Lazarsfeld se non attraverso una serie di esempi, tuttavia risulta evidente che i primi mirano a descrivere, mentre i secondi a prevedere. L'autore chiarisce che la distinzione in questione non è legata a caratteristiche proprie degli indicatori in se, ma «fa riferimento al posto che hanno gli indicatori nei processi ipotetici che mediano tra le osservazioni iniziali e l'immagine concettuale che viene sviluppata per organizzarli» (Lazarsfeld, 1958, tr. it. 1967, p. 194).

in quella riferita al periodo 2009-2010, in cui è soprattutto l'indicatore riferito alla rivista a determinare la classificazione finale, a un maggior numero di celle (6 anziché 4) viene assegnata la classe *undecided*. In sostanza gli indicatori relativi all'impatto/prestigio delle riviste vengono considerati indicatori meno validi di qualità della ricerca, rispetto al numero delle citazioni, e questo si riflette nella costruzione dell'indice sintetico.

La predisposizione di diverse matrici di sintesi è legata alla differenza tra le finestre temporali per il conteggio delle citazioni, eppure non risulta in grado di controllare completamente questo fattore. Le matrici di corrispondenza infatti sono due, mentre le finestre temporali di conteggio delle citazioni sono sette (una per ciascun anno incluso nell'esercizio di valutazione). In pratica la prima matrice (2004-2008) viene utilizzata su conteggi di citazioni effettuati su finestre temporali di 7, 6, 5, 4 e 3 anni, mentre la seconda (2009-2010) è riferita a conteggi su finestre di 2 o 1 anno.

Apparentemente, dato che il calcolo delle distribuzioni cumulative e dunque l'attribuzione delle classi avvengono per anno di pubblicazione, la differenza tra le finestre temporali non dovrebbe introdurre distorsioni, rendendo comparabili le classi ottenute su diverse finestre temporali, ma vi sono alcune questioni da considerare. Il tempo trascorso tra la pubblicazione e la prima citazione ricevuta, quello che intercorre tra questa e il raggiungimento del picco citazionale e quello in cui l'articolo continuerà ad essere citato (Garfield, 1979b) variano per campo di studi, genere del documento, genere di rivista e caratteristiche dell'autore come la notorietà o il prestigio (van Raan, 2005).

Ciò significa che la normalizzazione per campo di studi e anno di pubblicazione può essere resa inutile dalle differenze nelle finestre temporali; se il campo di studi e il genere dell'articolo non sono gli unici due fattori a influire sulla distribuzione delle citazioni, la procedura non risulta in grado di controllare le principali fonti di variazione. Ad esempio le posizioni nel ranking di articoli con lo stesso potenziale e le stesse caratteristiche, nella stessa *subject category*, potrebbero variare sensibilmente se osservati per periodi di diversa estensione. Possono essere individuate diverse soluzioni per ridurre questi rischi di distorsione, ma la più semplice è indubbiamente l'utilizzo di un'unica finestra temporale (*cfr*. § 6.1.2).

La procedura di sintesi appena presentata non tiene conto di una questione centrale: la definizione delle classi di merito va riferita alla classe finale, non alla classificazione dei prodotti sui due indicatori distinti. La riconduzione del numero di citazioni ricevute e dell'indice di impatto alle classi di merito e la sintesi per via tipologica delle classi non conducono necessariamente a una distribuzione della produzione scientifica mondiale con il 20% di prodotti eccellenti, il 20% di prodotti buoni, il 10% di prodotti accettabili ed il 50% di prodotti limitati. A seconda delle distribuzioni dei due indicatori, degli eventuali pareggi e soprattutto degli esiti della combinazione delle due classificazioni, potrebbe essere necessario modificare le soglie al fine di ottenere una distribuzione finale che rispetti la definizione delle classi di merito.

Le distribuzioni ottenute applicando gli algoritmi utilizzati nel corso della VQR non corrispondono alla distribuzione teorica prevista per le classi di merito[154]. Per questa ragione nel

[154] Il coordinatore dell'esercizio ha sottolineato che la calibratura degli algoritmi è problematica perché: «questa operazione è stata fatta a livello di area, è stata fatta un po' in fretta, è stato lasciato un margine di libertà eccessivo ai GEV per cui alla fine la probabilità di ottenere una valutazione eccellente per i vari GEV all'interno del database ISI o Scopus non era la stessa (Intervista Benedetto).

rapporto finale dell'Anvur un'intera appendice (Anvur, 2011a, Appendice A) è dedicata alla "calibrazione" degli algoritmi bibliometrici.

La calibrazione degli algoritmi bibliometrici ha comportato innanzitutto l'attribuzione di ciascuna delle 183 *subject categories* di WoS alle 14 Aree CUN[155], poi la riconduzione di tutti gli articoli presenti nel database alle quattro classi di merito, sulla base degli algoritmi utilizzati dal GEV corrispondente alla *Subject Category* dell'articolo. Gli articoli che nella procedura VQR sarebbero stati da inviare ad *informed review* (IR), cioè classificati in modo non coerente in base ai due indicatori (citazionale e di impatto), «sono stati ridistribuiti nelle quattro classi VQR in base alle distribuzioni, cella per cella, delle valutazioni peer ottenute dai prodotti *selezione 10%* di quel GEV» (Anvur, 2013a, Appendice 1, p. 1).

Le percentuali risultanti nelle quattro classi per ogni GEV sono state infine confrontate con quelle teoriche previste dal Bando e, sulla base di questo confronto, sono state calcolate le percentuali di correzione per la calibrazione (Tabella 34 e Tabella 35). Le distribuzioni delle classi di merito dell'universo degli articoli in WoS, derivanti dagli algoritmi messi a punto dai singoli GEV, sono risultate molto varie, spesso molto lontane dalle percentuali teoriche previste dal bando (Anvur, 2013a, Appendice 1). Le percentuali empiriche più vicine a quelle teoriche sono quelle delle Aree 5 e 6, rispettivamente Scienze Biologiche e Scienze Mediche.

E' possibile ottenere una misura sintetica della divergenza tra la classificazione empirica e la classificazione teorica utilizzando l'indice di accostamento chi quadrato[156]. Quest'indice assume valore 0 quando le differenze sono nulle, e anche se generalmente viene utilizzato per individuare la distribuzione teorica (normale, t di Student, chi quadrato, Poisson, binomiale, ecc.) che meglio si accosta alla distribuzione empirica (Di Ciaccio e Borra, 1996), in questo caso può fornire una misura sintetica della distanza tra la distribuzione delle classi definita dall'Anvur (20,20,10,50) e le distribuzioni empiriche nelle diverse Aree (Tabella 34).

Tabella 34 -Percentuali nelle quattro classi VQR dell'universo WoS, a seguito dell'applicazione degli algoritmi bibliometrici messi a punto dai singoli GEV (tabella A1.1, Anvur, 2013a, Appendice 1, p. 2)

	GEV01	GEV02	GEV03	GEV04	GEV05	GEV06	GEV07	GEV08	GEV09	GEV11
Eccellente	37,27%	24,59%	23,43%	23,54%	22,10%	21,95%	35,15%	37,92%	39,67%	36,34%
Buono	22,32%	27,50%	28,26%	24,59%	21,24%	21,04%	23,52%	23,29%	24,98%	12,46%
Accettabile	10,62%	19,82%	10,66%	11,64%	10,68%	12,57%	4,81%	5,27%	13,80%	12,32%
Limitato	29,78%	28,09%	37,64%	40,23%	45,97%	44,44%	36,52%	33,53%	21,55%	38,88%

Tabella 35 - Fattori di correzione delle percentuali risultanti dall'algoritmo di calibrazione (tabella A1.2, *ivi*)

	GEV01	GEV02	GEV03	GEV04	GEV05	GEV06	GEV07	GEV08	GEV09	GEV11
Eccellente	17,27%	4,59%	3,43%	3,54%	2,10%	1,95%	15,15%	17,92%	19,67%	16,34%
Buono	2,32%	7,50%	8,26%	4,59%	1,24%	1,04%	3,52%	3,29%	4,98%	-7,54%
Accettabile	0,62%	9,82%	0,66%	1,64%	0,68%	2,57%	-5,19%	-4,73%	3,80%	2,32%
Limitato	-20,22%	-21,91%	-12,36%	-9,77%	-4,03%	-5,56%	-13,48%	-16,47%	-28,45%	-11,12%

[155] L'attribuzione delle *subject categories* è stata effettuata «sulla base di una regola di maggioranza applicata alla distribuzione delle riviste ospitanti i prodotti VQR. Più precisamente, una SC è stata attribuita a una sola area se questa aveva più del 75% dei prodotti nelle riviste di quella SC, altrimenti alle due aree con il maggior numero di prodotti VQR pubblicati sulle riviste appartenenti a quella SC» (Anvur, 2013a, Appendice 1, p. 1).

[156] L'indice di accostamento chi quadrato è, sostanzialmente, una somma ponderata delle differenze quadratiche relative, con pesi pari alle frequenze teoriche: $\sum_{j=1}^{K} \frac{(n_j+n*_j)^2}{n*_j}$, dove n_j sono le frequenze empiriche, $n*_j$ le frequenze teoriche e j (da 1 a K) il numero delle classi (Di Ciaccio e Borra, 1996, pp. 189-191).

Tabella 36 – Indice di accostamento chi-quadrato e probabilità a una coda per Area[157]

	GEV01	GEV02	GEV03	GEV04	GEV05	GEV06	GEV07	GEV08	GEV09	GEV11
Indice di accostamento chi- quadrato	23,40	23,11	7,10	3,86	0,67	1,52	18,42	24,26	38,22	19,20
Probabilità a una coda di H_0	0,000	0,000	0,069	0,277	0,881	0,677	0,000	0,000	0,000	0,000

Il valore massimo di quest'indice, come quello del chi quadrato classico, dipende da una serie di fattori[158], dunque per valutare l'entità della differenza è necessario fare riferimento alla distribuzione di probabilità del chi-quadrato. L'ipotesi che la distribuzione dei prodotti nell'Area rispecchi la distribuzione teorica (H_0) può essere accettata senza problemi per le Aree 4, 5, 6 e con un margine di sicurezza inferiore per l'Area 3, ma va rifiutata per tutte le altre Aree (Tabella 36).

Dal punto di vista metodologico questa evidenza costituisce un problema centrale: le procedure poste in atto per classificare i prodotti producono un esito non coerente con il criterio quantitativo posto alla loro base. Ad esempio nell'Area 1 il 37,27% dei prodotti sono classificati come *eccellenti*, e solo il 29,78% come *limitati*: ciò vuol dire che l'eccellenza copre una quota più elevata dei prodotti rispetto a quella prevista dalla procedura di valutazione, mentre la quota di prodotti di qualità limitata risulta molto ridotta rispetto alle attese. Nell'Appendice A si legge che «secondo il DM e il Bando, la classe di "eccellente" viene attribuita a una pubblicazione che "si colloca nel 20% superiore della scala di valore condivisa dalla comunità scientifica internazionale", e analoghe definizioni valgono per le altre tre classi. In termini precisi dal punto di vista statistico, tale definizione significa che, estraendo a caso un articolo nell'insieme di tutti gli articoli pubblicati nel mondo in una certa area scientifica, la sua probabilità di essere eccellente secondo la definizione VQR è pari a 0,2.» (Anvur, 2013a, Appendice A, p. 1). Dunque, in altri termini, la probabilità che un prodotto venga classificato come eccellente risulta più alta del previsto, mentre quella che un prodotto sia valutato come limitato è più ristretta.

La mancata calibratura degli algoritmi conduce a una distorsione nella classificazione, verosimilmente in grado di spiegare, almeno in parte, l'elevata quota di prodotti eccellenti nelle Aree che hanno utilizzato la procedura bibliometrica. Come riportato nel rapporto finale: «una calibrazione accurata di questi algoritmi non è stata possibile prima dell'approvazione e pubblicazione dei criteri per mancanza di tempo e per l'indisponibilità delle basi di dati nei primi mesi del 2012» (Anvur, 2013a, p. 28), tuttavia la mancanza di una calibrazione, insieme all'utilizzo di algoritmi di attribuzione delle classi di merito differenti ha innanzitutto comportato rilevanti difformità negli esiti della valutazione rispetto alle loro premesse, inoltre ha reso sostanzialmente inconfrontabili i risultati ottenuti dalle strutture in Aree diverse.

Nel caso dei prodotti indicizzati su entrambi i database si pone un'ultima problematica, relativa a quale delle due valutazioni utilizzare. Si è già accennato e sarà argomentato più approfonditamente nel Capitolo 5, infatti, che le diverse basi di dati, in ragione delle differenze

[157] Si noti che non essendo stato reso noto il numero di prodotti classificati dalla procedura in ciascuna Area l'indice è stato calcolato non in base alle frequenze assolute ma a quelle relative (percentuali). Se da un lato dunque l'indice ottenuto non tiene conto della numerosità dei prodotti in ciascuna area, dall'altro questa numerosità non influenza il valore dell'indice, permettendo un confronto al netto delle dimensioni dell'Area.

[158] In questo caso il massimo empirico dell'indice, corrispondente alla classificazione empirica di tutti i prodotti nella classe teorica più esigua (Accettabile, 10%), è pari a 900.

della loro copertura, degli algoritmi di calcolo, degli indicatori e delle classificazioni tematiche, possono condurre a esiti diversi, assegnando diverse classi di merito a uno stesso prodotto. Nel rapporto finale di Area 3 è esposto chiaramente che: «nel caso di articoli sottoposti a valutazione bibliometrica utilizzando sia ISI WoS che Scopus che ottengano una valutazione diversa si adotterà la valutazione più favorevole» (Anvur, 2013d, GEV3, p. 94)[159].

La scelta di utilizzare sempre la valutazione migliore mira a non penalizzare in alcun modo i prodotti indicizzati su entrambi i database, dunque dal punto di vista pragmatico risulta un criterio estremamente condivisibile. Ciò nondimeno mette a rischio l'effettiva rispondenza dell'esito delle procedure con le soglie definite dal bando per le classi di merito, contribuendo a produrre il vantaggio delle Aree bibliometriche rispetto alle altre (la distorsione sistematica a favore delle valutazioni tramite analisi bibliometrica emerge chiaramente dal confronto con le valutazioni peer; Anvur, 2013a, Appendice B).

Conclusioni

L'analisi metodologica dell'operativizzazione del concetto di qualità della ricerca nella VQR ha evidenziato soprattutto la carenza di trasparenza, e a volte di coerenza, nei rapporti dell'Anvur. La traduzione delle dimensioni concettuali prima in indicatori e poi in variabili, e la ricomposizione delle informazioni rilevate in un indice sintetico risultano solo parzialmente rendicontate, e dunque solo parzialmente pubbliche, ripetibili e controllabili.

Le schede di rilevazione utilizzate per la valutazione dei prodotti nella procedura di peer review non sono state rese pubbliche in tutte le Aree. La formulazione delle modalità di risposta, e in particolare la presenza di più oggetti e la mancanza di mutua esclusività ed esaustività insieme alla scelta di proporre solo quattro alternative, mettono a rischio l'affidabilità delle valutazioni, rischiando principalmente di spingere i revisori verso le modalità mediane. I punteggi assegnati a ciascuna modalità, pur risultando coerenti dal punto di vista semantico con la formulazione degli item, presuppongono un assunto di equidistanza delle categorie ordinate non comprovato né argomenato. Queste caratteristiche tutto sommato non sembrano produrre ulteriori distorsioni sulle informazioni rilevate, anche se sarebbe stato forse preferibile non rendere visibili ai revisori i punteggi corrispondenti a ciascuna modalità di risposta.

[159] Il criterio, confermato sia dai Coordinatori sub-GEV che dal Presidente di Area 3, era esteso anche ai casi in cui i prodotti fossero stati valutati sia tramite peer review che tramite analisi bibliometrica (il campione estratto casualmente per il confronto sperimentale, cfr. Anvur, 2013d, GEV3). In proposito la procedura prevedeva l'apertura di un gruppo di consenso (Anvur, 2013d, GEV3, p. 94), che tuttavia ha deciso quasi sempre per la valutazione migliore: «la filosofia era sempre quella di fare le scelte più favorevoli… per il prodotto» (Intervista Torsi); «ci sono stati dei casi ovviamente dove c'era coincidenza tra l'assegnazione e quindi era molto facile; casi in cui c'era una differenza di una unità, cioè classificato in A bibliometrico, B dai revisori, allora il gruppo di consenso faceva una breve discussione, in genere si optava per la situazione più favorevole, quindi diciamo per quella bibliometrica che spesso era più favorevole di quella peer review; e poi c'erano dei casi in cui la discrepanza era di due soglie o addirittura, magari in qualche caso di tre, allora lì caso per caso si è fatta una valutazione comparata tra quello che era l'indice bibliometrico e quello che era la valutazione del revisore, perché erano così diversi e in genere si è trovato un valore intermedio tra le due valutazioni» (Intervista Pacchioni).

La sintesi delle valutazioni, con l'attribuzione di una classe di merito al prodotto, pur avvenendo con un'operazione matematica, risulta sostanzialmente basata su considerazioni di ordine semantico, che però tanto quanto la procedura di sintesi in sé, non vengono esposte nel rapporto finale. Neppure questa fase è esente da criticità legate da un lato al rapporto degli indicatori con il concetto e delle loro relazioni reciproche, dall'altro al tipo di variabili sintetizzate. Infine la riconduzione delle classi assegnate dai singoli *referee* alla classe di merito finale è uno dei punti più oscuri della procedura di valutazione dei prodotti nella VQR. Le informazioni disponibili su questa fase sono scarse e frammentarie, non di rado risultano incoerenti. La mancanza di trasparenza è uno dei nodi cruciali nel dibattito sulla VQR, e non a caso.

La scelta degli indicatori bibliometrici viene scarsamente argomentata in relazione al concetto di qualità della ricerca alla base dell'esercizio di valutazione e le debolezze degli indicatori utilizzati non vengono presentate né discusse. Inoltre vi è una totale assenza di riferimenti alle differenze, semantiche e tecniche, tra l'*impact factor* e il SJR. Il legame tra gli indicatori selezionati ed il concetto di qualità della ricerca andrebbe ridiscusso tanto dal punto di vista semantico quanto dal punto di vista tecnico. Non sono poche infatti le questioni aperte circa l'affidabilità degli indicatori impiegati, in particolare considerando l'utilizzo di diversi indicatori, costruiti in base a differenti database.

La procedura di riconduzione degli indicatori grezzi alle classi di merito, pur risolvendo gran parte delle questioni legate al confronto tra discipline differenti e prodotti diversi per tipologia e anno di pubblicazione, non risulta del tutto chiara nei rapporti. Nuovamente inoltre va sottolineato che i problemi procedurali, tecnici, classificatori e concettuali legati all'utilizzo di due diversi database non vengono discussi o affrontati.

Infine la procedura di sintesi controlla solo parzialmente le differenze tra i prodotti in termini di anni trascorsi dalla pubblicazione e non è in grado di rispettare la definizione quantitativa delle classi di merito prevista nel decreto ministeriale. Le quote di prodotti previste per le quattro classi di merito vengono infatti riferite nella procedura alla distribuzione singola dei due indicatori utilizzati e non alla loro distribuzione congiunta.

L'utilizzo degli strumenti di analisi bibliometrica nel corso della VQR è stato evidentemente preceduto da una riflessione mirata circa le possibilità di confronto degli indicatori per discipline differenti e prodotti diversi per tipologia e anno di pubblicazione, non accompagnata però dalla considerazione delle complicazioni derivanti dall'impiego di più di un database.

Dal punto di vista metodologico la carenza di attenzione nella costruzione del dato e nella sua rendicontazione costituisce un rischio capitale, non solo in riferimento all'affidabilità della definizione operativa, ma anche i riferimento alla solidità dei presupposti dell'indagine: «quando il linguaggio non è semanticamente preciso, i processi di operazionalizzazione sono carenti, deboli quelli osservativi e di analisi, le "pretese" massimizzano il loro carattere di "invenzione", minimizzano quello attinente alla loro giustificazione logica. I "punti di vista" allora prevalgono, la produzione scientifica si lega facilmente ai trends culturali, alle posizioni ideologiche, alle condizioni sociali, e ne segue il destino» (Bruschi, 1996 p. 169).

Capitolo 5
La rilevazione della qualità

Introduzione

Il capitolo affronta la questione della rilevazione dei dati nella valutazione dei prodotti della ricerca nel corso della VQR 2004/2010. Dopo aver analizzato le definizioni operative della qualità della ricerca messe a punto nell'Area 14 per la procedura di valutazione in peer review e nell'Area 3 per la procedura di valutazione diretta tramite analisi bibliometrica, sono messe a fuoco le problematiche relative alla selezione e alle caratteristiche delle fonti delle informazioni: i revisori peer e i database bibliometrici. Inevitabilmente il discorso sarà nuovamente riferito alle procedure applicate dai singoli GEV, e dunque principalmente alle due Aree selezionate come casi-studio.

In riferimento alla procedura di peer review saranno prese in considerazione la selezione dei pari, l'assegnazione dei prodotti da valutare e le possibilità di reazione all'oggetto della valutazione. A proposito della procedura di valutazione diretta tramite analisi bibliometrica, il discorso riguarderà invece soprattutto la scelta dei database e le loro caratteristiche. In sintesi, come già affermato nel Capitolo 2, ci si occuperà, in entrambi i casi, di quanto l'insieme delle informazioni (valutazioni) rilevate possa dirsi non distorto sulla base di *chi* (o *cosa*) le ha fornite, mentre fino a questo punto si è messo a fuoco il *come* le informazioni sono state richieste ed elaborate. Come evidenziato da Fasanella infatti «al fine di validare la valutazione non si può prescindere dalla base di dati su cui essa poggia» (2014, p. 132).

5.1 La rilevazione nella procedura di valutazione tramite peer review nella VQR

La rilevazione dei dati circa la qualità dei prodotti della ricerca nella procedura di peer review consta essenzialmente nella selezione dei revisori, nell'assegnazione dei prodotti ai revisori e nella registrazione delle loro valutazioni di ciascun prodotto. E' a queste tre fasi essenziali che si farà riferimento nell'analisi della rilevazione dei dati.

Nel corso della VQR la selezione dei revisori ha previsto innanzitutto la costruzione di un Albo di revisori Anvur, suddiviso per GEV. Questo albo è stato costruito sulla base dell'albo CINECA dei revisori PRIN e FIRB, selezionando però i revisori tra i nominativi registrati in base a criteri di merito scientifico. Nel rapporto finale dell'Agenzia i criteri indicati sono essenzialmente bibliometrici: h *Index*, numero di citazioni e produzione scientifica recente (Anvur 2013a, p. 24), tuttavia viene segnalato altresì che «ovviamente, la scelta dei criteri è stata modulata dai vari GEV a seconda della disponibilità o meno di informazioni di natura bibliometrica» (*ibidem*). La costruzione di un nuovo

albo a partire dall'albo CINECA si è resa necessaria stante che le credenziali scientifiche dei revisori iscritti all'albo non erano mai state sottoposte a valutazione e che il numero di revisori stranieri risultava limitato (*ibidem*). In una seconda fase l'albo è stato integrato con l'inserimento di un numero elevato di esperti selezionati sulla base degli stessi criteri e interpellati individualmente al fine di valutarne la disponibilità a partecipare alla VQR. In questa stessa fase l'Anvur ha pubblicato un modulo di auto-candidatura, compilabile da coloro che, non essendo già presenti nell'Albo dei revisori, intendessero contribuire al processo di valutazione. Si è così pervenuti all'albo iniziale Anvur-VQR, comprensivo di oltre 16.000 nominativi (*ibidem*) [160].

La selezione dei revisori inoltre è proseguita anche durante la fase di valutazione «per coinvolgere competenze non coperte dalle liste definite fino a quel momento e resesi necessarie per la valutazione di prodotti specifici» (*ibidem*).

Nessun dato circa la rispondenza dei revisori ai criteri definiti dai GEV è stato tuttavia reso noto, con la sola eccezione dell'affiliazione italiana o straniera. Nel complesso l'esercizio di valutazione ha coinvolto 14.770 revisori attivi, quasi un terzo dei quali di affiliazione straniera (il 27,1%, Anvur, 2013b, p. 25, *cfr.* Tabella 37).

Tabella 37 - Numero di revisori per Area e affiliazione italiana o straniera (adattamento Tab. 3.1 Anvur, 2013b, p. 25)

Area		Revisori con affiliazione italiana		Revisori con affiliazione straniera		Totale
		N	%	N	%	
1	Scienze matematiche e informatiche	166	29,4%	398	70,6%	564
2	Scienze fisiche	828	66,9%	409	33,1%	1.237
3	Scienze chimiche	528	81,5%	120	18,5%	648
4	Scienze della Terra	245	86,0%	40	14,0%	285
5	Scienze biologiche	849	71,3%	341	28,7%	1.190
6	Scienze mediche	1.374	81,2%	318	18,8%	1.692
7	Scienze agrarie e veterinarie	510	77,9%	145	22,1%	655
8	Ingegneria civile ed Architettura	544	62,3%	329	37,7%	873
9	Ingegneria industriale e dell'informazione	231	18,7%	1.005	81,3%	1.236
10	Scienze dell'antichità, filologico-letterarie e storico-artistiche	1.763	73,1%	649	26,9%	2.412
11	Scienze storiche, filosofiche, pedagogiche e psicologiche	1.076	74,6%	366	25,4%	1.442
12	Scienze giuridiche	1.285	93,9%	83	6,1%	1.368
13	Scienze economiche e statistiche	344	56,4%	266	43,6%	610
14	Scienze politiche e sociali	407	72,9%	151	27,1%	558
Totale		*10.150*	*68,7%*	*4.620*	*31,3%*	*14.770*

In sole due Aree il numero di revisori con affiliazione straniera supera quello dei revisori con affiliazione italiana, ma di misura: nell'Area 1 i *referee* stranieri sono il 70,6%, in Area 9 l'81,3% (Tabella 37).

[160] La sintesi del coordinatore dell'esercizio è: «noi abbiamo costruito un albo di revisori, in parte utilizzando quello che era già disponibile del Ministero, che però non aveva nessun requisito qualitativo della scelta, era semplicemente di volontari, per lo più italiani. Quindi ciascun GEV ha stilato un elenco di nomi, a partire da quelli che ha trovato sul sito del MIUR ma in realtà aggiungendone molti ed eliminandone moltissimi. L'insieme dei revisori, che sono circa 15.000, che noi abbiamo utilizzato costituiscono un database nostro, a questo punto, per il quale abbiamo fatto invece valutazioni ex ante della qualità, tipo indicatori bibliometrici per quelli che li avevano, tipo reputazione» (Intervista Benedetto).

La fase successiva della rilevazione della qualità prevedeva l'assegnazione dei prodotti ai revisori da parte dei membri del GEV. Nel rapporto finale dell'Anvur viene però posto in evidenza che «parte delle revisioni peer sono state effettuate da membri GEV, con le stesse procedure delle revisioni esterne. Tali revisioni interne hanno riguardato soprattutto i prodotti che la valutazione bibliometrica aveva assegnato alla classe IR (cioè *informed review*) che, quindi, richiedevano una valutazione peer. Complessivamente, la percentuale di revisioni peer effettuate direttamente all'interno dei GEV è stato contenuto e pari al 16,5%, percentuale prossima a quella dei prodotti assegnati dalla valutazione bibliometrica alla classe IR» (Anvur, 2013a, p. 25).

La procedura di rilevazione non è stata dunque rispettata per tutti i prodotti; il fatto che le revisioni da parte dei membri del GEV riguardino soprattutto i prodotti classificati come "undecided" dalla procedura bibliometrica, non costituisce un attenuante, dato che anche per questi prodotti erano previste due revisioni da parte di due diversi *referee* esterni al GEV. Inoltre si è già evidenziato in riferimento all'Area 14 come i dati suggeriscano un uso delle revisioni interne non limitato alle Aree bibliometriche, ma esteso ai prodotti "*undecided*" anche nelle Aree non bibliometriche.

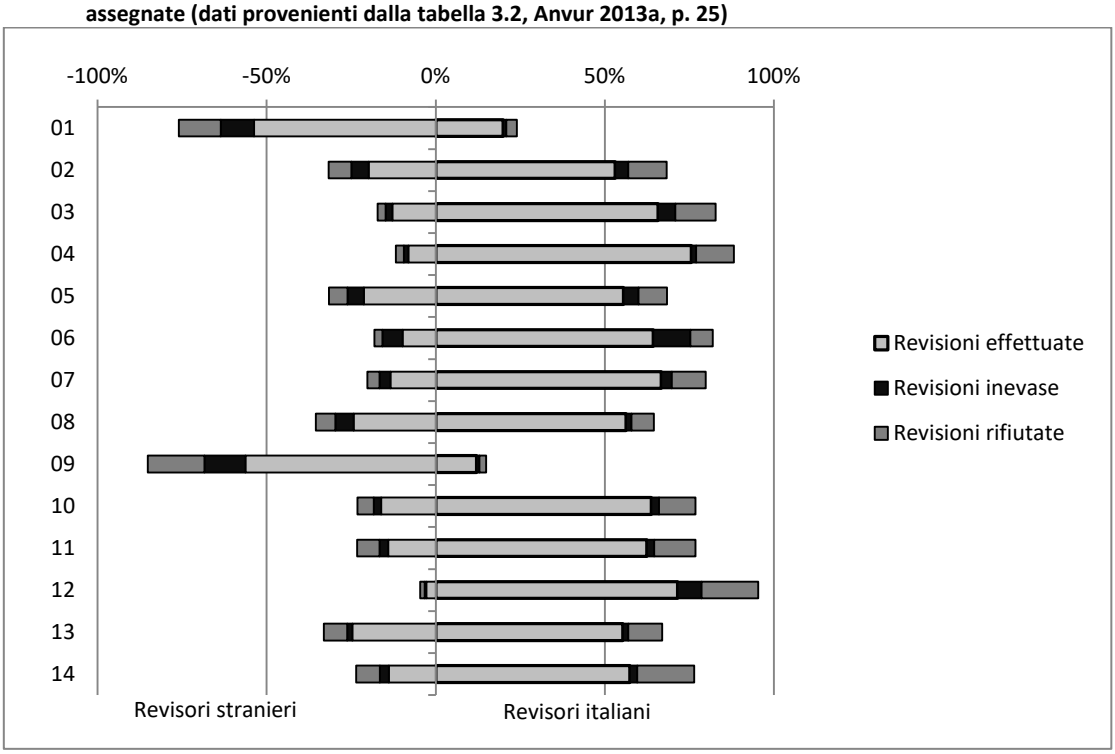
La mancata assegnazione del prodotto ai revisori non è l'unica possibile deviazione dalla procedura stabilita; i revisori avevano infatti la possibilità di rifiutare la presa in carico di un prodotto, per ragioni legate alla lingua, al campo di studi, a conflitti di interesse o alla semplice mancanza di tempo, oppure di non effettuarla, lasciandola inevasa pur avendo preso in carico il prodotto. L'Anvur, segnalando che «in qualche caso, per il ritardo nella consegna della valutazione da parte di alcuni revisori, circostanza che ha suggerito l'invio a un terzo revisore, il numero di revisioni è stato superiore a 2» (Anvur, 2013a, p. 25)[161], riporta i dati relativi alle revisioni assegnate, effettuate, inevase e rifiutate (*ibidem*, tabella 3.2).

Il Grafico 3 rappresenta le quote sul totale delle revisioni assegnate, delle revisioni effettuate, inevase e rifiutate, per Area e affiliazione dei revisori (per semplificare la rappresentazione grafica le quote relative ai revisori stranieri sono riportate con il segno negativo, in modo da poter osservare sulla sinistra le revisioni assegnate a *referee* stranieri e sulla destra quelle assegnate a *referee* italiani).

Le revisioni inevase e quelle rifiutate da parte di revisori stranieri risultano più frequenti nelle due Aree con il maggior numero di revisori con affiliazione estera, l'Area 1 e l'Area 9, ma tutto sommato le quote risultano proporzionali a quelle nelle altre Aree. Le revisioni rifiutate da revisori italiani sono invece più frequenti in Area 12 e Area 14, quelle inevase presentano quote relativamente più elevate in Area 6 e in Area 12.

[161] Il coordinatore dell'esercizio ha sottolineato che: «per ciascun libro o articolo che fosse, venivano individuati inizialmente due revisori, se poi uno dei due non rispondeva in tempo si passava a un altro… per cui alla fine tutti hanno avuto almeno due revisioni, ma parecchi ne hanno avuti anche tre o quattro perché la persona non rispondeva, allora se ne trovava un terzo, ma a quel punto venivano effettuate tutte e due. Noi abbiamo utilizzato tutte quelle di cui disponevamo» (Intervista Benedetto). In altri termini nel caso in cui a seguito di un ritardo da parte di uno dei revisori il GEV riassegnasse il prodotto e il revisore in questione inviasse comunque la scheda tutte le revisioni disponibili venivano utilizzate al fine dell'assegnazione della classe di merito finale, sulla base dello schema predefinito dall'Agenzia (Anvur, 2014).

Grafico 3 – Revisioni effettuate, inevase e rifiutate per Area e affiliazione dei revisori, % sul totale delle revisioni assegnate (dati provenienti dalla tabella 3.2, Anvur 2013a, p. 25)



La mancata effettuazione di una revisione rende necessaria la riassegnazione del prodotto, dunque la sua valutazione da parte di un revisore altro da quello inizialmente selezionato. Naturalmente non è possibile evitare questo genere di imprevisti, nondimeno sarebbe opportuno un approfondimento delle loro conseguenze sulla procedura di valutazione, anche e soprattutto perché la loro incidenza non è omogenea tra le Aree e può dipendere da diversi fattori. Mettendo meglio a fuoco i dati disponibili (Tabella 38), è possibile ad esempio notare la maggiore quota di rifiuti tra le revisioni assegnate a *referee* stranieri (nel complesso pari a circa il 20%, contro il 14% dei rifiuti da parte di revisori italiani). Una spiegazione semplice potrebbe essere individuata nella lingua di pubblicazione, considerando che le quote di rifiuti da parte di revisori stranieri sono più elevate nelle Aree con una maggior quota di prodotti in lingua italiana: le Aree 10, 11, 12 e 14 (Anvur, 2013a, figura 2.3, p. 17).

Nel complesso l'Area con la maggior quota di rifiuti, sia da parte di revisori italiani che da parte di revisori stranieri, è l'Area 14, mentre la maggior quota di revisioni inevase, sia da parte di revisori italiani che da parte di revisori stranieri, è l'Area 6 (Tabella 38). L'Area 9 presenta la minor quota di revisioni effettuate in rapporto al totale di quelle assegnate.

Tabella 38 – Numero di revisioni assegnate e quote di revisioni effettuate, inevase e rifiutate per Area e affiliazione dei revisori (adattamento della tabella 3.2, Anvur 2013a, p. 25)

Area	Revisori italiani				Revisori Stranieri				Totale revisori			
	Revisioni assegnate (v.a.)	Revisioni effettuate (%)	Revisioni inevase (%)	Revisioni rifiutate (%)	Revisioni assegnate (v.a.)	Revisioni effettuate (%)	Revisioni inevase (%)	Revisioni rifiutate (%)	Revisioni assegnate (v.a.)	Revisioni effettuate (%)	Revisioni inevase (%)	Revisioni rifiutate (%)
01	1.362	83,0	3,9	13,1	4.317	70,7	12,9	16,3	5.679	73,7	10,8	15,5
02	8.799	77,7	5,6	16,6	4.086	62,5	16,3	21,2	12.885	72,9	9,0	18,1
03	2.556	79,4	6,1	14,5	531	74,8	11,3	13,9	3.087	78,6	7,0	14,4
04	1.429	85,7	1,7	12,7	191	68,6	11,0	20,4	1.620	83,6	2,8	13,6
05	5.183	81,2	6,5	12,3	2.392	67,4	15,2	17,4	7.575	76,8	9,2	13,9
06	20.296	78,6	13,3	8,1	4.483	54,1	32,9	13,0	24.779	74,2	16,9	9,0
07	8.374	83,5	3,9	12,6	2.119	66,3	16,2	17,6	10.493	80,0	6,4	13,6
08	11.202	87,3	2,4	10,3	6.158	68,6	15,0	16,4	17.360	80,6	6,9	12,5
09	2.705	81,0	5,9	13,1	15.480	66,1	14,2	19,7	18.185	68,3	13,0	18,7
10	26.156	83,0	3,0	14,0	7.902	69,8	9,2	21,0	34.058	80,0	4,4	15,6
11	21.445	81,3	2,9	15,8	6.481	60,6	10,9	28,5	27.926	76,5	4,8	18,8
12	27.566	75,0	7,4	17,6	1.324	60,3	11,5	28,2	28.890	74,3	7,5	18,1
13	8.928	82,7	2,3	15,0	4.408	74,7	4,4	20,9	13.336	80,1	3,0	16,9
14	9.138	75,0	2,8	22,1	2.811	58,8	11,3	29,9	11.949	71,2	4,8	24,0
Totale	**155.139**	**80,2**	**5,4**	**14,4**	**62.683**	**65,8**	**13,9**	**20,3**	**217.822**	**76,0**	**7,9**	**16,1**

La lingua di pubblicazione non è l'unico fattore in grado di incidere sull'accettazione o sul rifiuto di un prodotto, basti pensare al livello di specializzazione dei contenuti o a caratteristiche dell'autore come il ruolo accademico o l'istituzione di afferenza. Inoltre ciascuna di queste caratteristiche del prodotto potrebbe incidere diversamente sulla decisione di differenti revisori.

L'attenzione ai requisiti di trasparenza e pubblicità della procedura di valutazione con riferimento a questa fase è indubbiamente legata alle possibili distorsioni connesse alle caratteristiche di valutatori e prodotti. Si pensi al campo di studi, all'età, al genere, al prestigio e al ruolo accademico, in riferimento alle quali in letteratura sono note diverse distorsioni possibili (la questione è approfondita nel § 5.1.2; si vedano ad esempio: Crane, 1967; Zuckerman e Merton, 1973; Mahoney, 1977; Wennerâs e Wold, 1997; per una rassegna sul referaggio nelle riviste: Campanario 1998a e 1998b; per una rassegna più generale e recente: Lee *et al.* 2013).

Nonostante sia stata pubblicata la lista dei revisori VQR che hanno dato l'assenso alla pubblicazione del proprio nominativo (Anvur, 2014a), l'unico dato disponibile sulle loro caratteristiche è l'affiliazione italiana o straniera. L'elenco infatti contiene esclusivamente i nominativi, senza riportare né l'ambito disciplinare, né l'Area di riferimento, tantomeno la struttura di appartenenza o il ruolo accademico dei revisori.

Volendo escludere un livello di dettaglio che riportasse queste informazioni per ciascun revisore, nel caso della VQR sarebbe stato comunque possibile pubblicare l'incidenza di alcune caratteristiche tra i revisori per Area, ad esempio il genere, l'età, il ruolo accademico, il settore scientifico disciplinare di afferenza, il dipartimento (o anche solo l'Ateneo o l'Ente) di appartenenza, senza dimenticare i dati relativi alla produzione scientifica che avrebbero potuto dare un'idea della rispondenza dei revisori ai criteri di selezione dei GEV. Qualcosa di simile è stato fatto con

riferimento alle caratteristiche dei membri dei 14 GEV, seppure limitatamente all'affiliazione, al genere e, con riferimento agli EV di affiliazione italiana, all'Area geografica (Anvur, 2013a, p. 19, tabella 2.12; Tabella 39). I tre criteri principali erano: «(1) qualità scientifica (tenendo conto del merito scientifico, delle sedi di pubblicazione, del numero delle citazioni, dell'impatto della ricerca nella comunità internazionale e di eventuali premi di ricerca o altri riconoscimenti); (2) continuità della produzione scientifica negli ultimi 5 anni; (3) esperienza in attività di valutazione a livello nazionale e internazionale» (Anvur, 2013a, p. 18), cui si aggiungevano alcune condizioni: la copertura delle linee culturali e di ricerca all'interno delle Aree; almeno il 20% di membri GEV docenti in università straniere; l'attenzione alla distribuzione di genere; per i membri di affiliazione italiana, ove possibile, un'equa distribuzione di sede e un'equa distribuzione geogafica. E' evidente che non tutte le condizioni siano state pienamente rispettate (Tabella 39), ma qui vale la pena soprattutto di notare come i dati pubblicati si riferiscano solo a tre delle condizioni poste dall'Agenzia, e che dunque, neppure con riferimento ai GEV, è stata resa nota la rispondenza ai principali criteri di selezione previsti, o alla copertura delle linee culturali e di ricerca.

Tabella 39 – Distribuzione dei membri dei GEV per affiliazione, genere ed area geografica (adattamento della tabella 2.12; Anvur, 2013a, p. 19)

Affiliazione	**n.**	**%**
Italiana	360	80
Estera	90	20
Totale	450	100
Genere	**n.**	**%**
Donne	106	23,6
Uomini	344	76,4
Totale	450	100
Area geografica (affiliazione italiana)	**n.**	**%**
Nord	163	45,3
Centro	113	31,4
Sud	84	23,3
Totale	360	100

Una procedura realmente pubblica e controllabile avrebbe richiesto una maggiore trasparenza. Ad esempio, oltre alla pubblicazione delle distribuzioni di alcune caratteristiche dei revisori per Area, sarebbe stata estremamente rilevante la divulgazione di informazioni sintetiche circa l'incidenza delle caratteristiche rilevanti dei revisori in relazione alle caratteristiche dei prodotti valutati (ad esempio tipo di pubblicazione, anno di pubblicazione, lingua) e dei loro autori (di nuovo genere, età, ruolo accademico, affiliazione, ecc.).

5.1.1 I rilevatori della qualità nell'Area 14

Nell'Area 14 la procedura di invito/selezione dei revisori è iniziata nel Marzo del 2012, ed è rimasta aperta per molti mesi: «sia perché il Consiglio Direttivo dell'Anvur ha deciso di aprire una fase per le auto-candidature, sia per permettere ai membri GEV di sostituire i numerosi *referee* che hanno risposto all'invio dei prodotti con un rifiuto» (Anvur, 2013d, GEV14, p. 10).

Nel documento relativo ai criteri di Area si legge: «il GEV intende coinvolgere preferibilmente revisori esterni con un profilo di ricerca internazionale, un curriculum di alto profilo, testimoniato, in particolare negli ultimi anni, da un elevato numero di pubblicazioni nelle sedi di riferimento della comunità scientifica internazionale del settore, un significativo numero di citazioni e la necessaria competenza nella specifica area di valutazione» (Anvur 2013d, GEV14, p. 73).

I requisiti previsti per i revisori erano sostanzialmente:

(1) avere svolto attività scientifica a livello internazionale;
(2) possedere un curriculum di alto profilo (sufficiente numero di pubblicazioni, recenti, citate, adeguatamente collocate dal punto di vista editoriale);
(3) possedere la necessaria competenza nella specifica area della valutazione.

La definizione dei requisiti è dunque abbastanza lasca: date le caratteristiche delle discipline di Area non era infatti pensabile una adozione esclusiva o vincolante di criteri scientometrici per la selezione dei revisori[162]. Inoltre va tenuto conto del fatto che la partecipazione all'esercizio di valutazione come revisori era su base essenzialmente volontaria: anche nel caso ideale in cui tutti i profili rispondenti ai requisiti previsti fossero stati individuati e iscritti all'albo, i revisori sarebbero comunque risultati auto selezionati sulla base della loro disponibilità[163].

Sono stati selezionati inizialmente 443 revisori di Area 14, ma il numero totale di revisori coinvolti è cresciuto fino a 558, per due ragioni:

- la necessità di coinvolgere *referee* afferenti a diverse Aree (in particolare di Area 11) per la loro competenza rispetto ai prodotti;
- la necessità di allargare il numero di revisori a seguito di rifiuti e valutazioni inevase (Anvur, 2013d, GEV14, p.26).

L'insieme finale di revisori impiegati non corrisponde dunque a quello inizialmente identificato sulla sola base dei criteri di selezione prestabiliti. Nell'Area delle Scienze Politiche e Sociali, così come nelle altre Aree, non si procede a un confronto fra il profilo scientifico atteso in base ai vincoli iniziali, posti dall'Anvur e dal GEV 14, e il profilo scientifico effettivo dei revisori selezionati ai fini della valutazione. L'unico dato disponibile circa le caratteristiche dei revisori è, di nuovo, quello relativo all'affiliazione nazionale/straniera: 407 *referee* (il 72,9%) sono di affiliazione italiana, 151 (27,1%) di

[162] Le parole del presidente GEV sulle procedure e l'esito della selezione sono di estremo interesse: «per aumentare la qualità noi abbiamo cercato di avere dei *referee* di qualità il più possibile alta […] Come abbiamo ottenuto questo: attraverso vari tentativi, diciamo, e varie procedure. Essendo un settore non bibliometrico ci sono anche li, come lei sa, problemi molto chiari, perché l'h-index non funziona, però non si può prescindere totalmente dall'h-index […] allora abbiamo considerato l'h-index, abbiamo considerato la reputazione, abbiamo considerato il curriculum. Su questa base abbiamo fatto una prima selezione dei *referee*, molto ristretta, con *referee* che hanno h-index, reputazione e, diciamo, tipo di lavori (intesi come lavori pubblicati su sedi editoriali di prestigio, riviste di classe A con impact factor elevato, case editrici internazionali e nazionali note, ecc.) e abbiamo scelto questo primo gruppo. Il problema è che questo primo gruppo di *referee* si è dimostrato assolutamente insufficiente, quantitativamente, allora abbiamo dovuto allargare, e anche usare maglie un po' più larghe» (Intervista Colozzi).

[163] La volontarietà del compito è uno dei problemi fondamentali nell'opinione della professoressa Bazzicalupo: «questo è uno dei punti deboli del sistema: i valutatori sono volontari. Questo è complicato. Primo perché nonostante gli appelli, almeno io personalmente penso anche gli altri abbiano fatto, a rendersi disponibili a fare i valutatori, pochissimi hanno aderito perché è un lavoro abbastanza gravoso, noioso e poi forse non avevano neanche capito fino in fondo quanto era importante. Allora si è avuto un *pool* di valutatori molto ristretto rispetto all'enorme quantità di prodotti. Ovviamente questo significa attribuire allo stesso valutatore diversi prodotti, alcuni dei quali diversi, e questo crea problemi» (Intervista Bazzicalupo).

affiliazione straniera (*cfr.* Anvur, 2013d, GEV14, Tab. 2.12, p. 29). Nel rapporto si legge che «la necessità di allargare in maniera consistente la platea dei revisori ha alterato la proporzione tra italiani e stranieri che inizialmente era molto meno squilibrata (seppur sempre sbilanciata a favore dei primi)» (*ivi* p. 26).

La procedura prevedeva dunque una serie di *standard* per la selezione dei revisori, ma non è possibile alla luce dei dati pubblicati valutare la rispondenza delle modalità di selezione impiegate agli standard prestabiliti, perlomeno non in riferimento al loro esito[164]. E' invece possibile effettuare alcune considerazioni in relazione ai requisiti di trasparenza e pubblicità della procedura stessa.

La selezione dei revisori è scarsamente discussa tanto nel documento relativo ai criteri quanto nel rapporto finale e i dati pubblicati circa i revisori coinvolti non fanno riferimento in alcun modo al loro profilo disciplinare o scientifico. La stessa definizione dei criteri di selezione risulta estremamente vaga, non fornisce dei veri e propri standard né dei requisiti minimi. Da un lato questa definizione risulta funzionale alle caratteristiche delle discipline e alle differenze tra i settori disciplinari interni all'Area, dall'altro offre il fianco a una critica di eccessiva discrezionalità. La selezione *in itinere* dei revisori non fa che accentuare queste criticità evidenziando una sorta di abbassamento progressivo degli standard in relazione alle necessità pratiche e alle tempistiche dell'esercizio di valutazione.

Riguardo la selezione dei revisori, è il caso di ricordare, andrebbero prese in considerazione una serie di possibili distorsioni legate alle caratteristiche dei revisori stessi, in particolare il campo di studi, il prestigio e il ruolo accademico (Lee *et al.* 2013) che potrebbero intervenire in diversa misura anche in relazione alle caratteristiche dei prodotti da valutare o dei loro autori. Qui si intende come distorsione (bias) qualsiasi caratteristica, cognitiva o attitudinale, del valutatore che potrebbe interferire con una valutazione obiettiva (Shatz, 2004).

Innanzitutto andrebbe considerata con maggiore attenzione la nota distintiva dei revisori *peer*, cioè la loro appartenenza alla stessa comunità degli autori dei prodotti da valutare, tenendo conto delle peculiarità disciplinari dell'Area delle Scienze Politiche e Sociali. L'individuazione della comunità dei pari non è affatto scontata nelle scienze sociali. La maggior parte delle discipline che afferiscono a quest'area sono caratterizzate da uno stato sostanzialmente pre (o multi) paradigmatico (Kuhn, 1962; 1977) dunque le comunità scientifiche e lo stesso consenso scientifico presentano caratteristiche del tutto peculiari (Knorr, 1975; Moody 2004) rispetto a campi disciplinari caratterizzati da una *normalità* (kuhnianamente intesa) di procedure e teorie e, dunque, da un consenso più uniforme e condiviso. Se il carico di conoscenze e competenze dei revisori è inevitabilmente accompagnato da un insieme di opinioni e aspettative soggettive (Kassirer e Campion, 1994; Hojat *et al.* 2003), indipendentemente dal campo disciplinare di riferimento, a

[164] L'opinione del Presidente GEV non evidenzia una forte rispondenza dell'esito ai criteri della selezione: «certamente se lei intende per qualità dei dati un criterio rigoroso per cui hanno fatto il referaggio solo i migliori studiosi di scienza politica e di sociologia disponibili a livello mondiale, no. Secondo me la maggior parte di quelli che hanno queste caratteristiche c'erano, quelli che non c'erano è perché non hanno accettato di farlo. Avevamo anche il problema dei rifiuti, abbiamo chiesto ad alcuni, molto prestigiosi, non hanno accettato. Chiaramente a quel punto non puoi obbligarli, soprattutto se non sono italiani. Quindi abbiamo dovuto sostituire alcuni dei migliori, che non ci sono stati, con persone di livello meno eccellente, diciamo, e poi abbiamo dovuto allargare. Alla fine il gruppo dei *referee* nella sua complessità, forse rappresentava di più la media del livello di produzione dell'Area 14 che non solo l'eccellenza. E questo può essere stato magari un primo fattore di distorsione. D'altra parte l'alternativa che avevamo era di gestire più della metà dei prodotti tutti all'interno del GEV, con rischi molto più elevati, ovviamente» (Intervista Colozzi).

questo insieme di opinioni e aspettative soggettive nell'ambito delle scienze sociali si aggiungono opinioni e aspettative "di scuola" condivise da porzioni della comunità scientifica più ampia ma non necessariamente dalla sua totalità (Mahoney, 1977; Cole 2000). L'appartenenza dei revisori peer alla stessa comunità degli autori, nelle scienze sociali, può essere definita a diversi livelli (tematico, metodologico, teorico) e non dovrebbe essere data per scontata né ritenuta a-problematica, soprattutto con riferimento alla valutazione della ricerca. Le discrepanze tra i giudizi nella peer review possono infatti non essere dovuti a un effettivo disaccordo dei pari circa la qualità del prodotto, ma piuttosto spiegate da diversità nelle posizioni teoriche o metodologiche, nei criteri e nelle aree di competenza tra revisori (Eckberg, 1991; Kostoff, 1995; Bornmann, 2011)[165].

Una selezione dei revisori che tenga conto delle diverse scuole teoriche o metodologiche potrebbe da un lato accrescere il grado di informazione sul prodotto, offrendone una valutazione da diverse prospettive, dall'altro condurre a un basso grado di accordo tra i *referee* (Marsh e Ball, 1991). Secondo alcuni la diversità di prospettive e criteri tra i revisori è utile al processo decisionale (nell'editoria, ma il discorso regge anche in ambito valutativo, soprattutto con riferimento alla valutazione dei progetti), tanto che è proprio su questa base che dovrebbe essere effettuata la scelta dei revisori (tra gli altri Stricker, 1991).

La mancanza di argomentazioni o dati su queste ed altre caratteristiche dei *referee* non permette di avanzare osservazioni più specifiche con riferimento alla VQR, né del resto sarebbe stato possibile valutare l'adeguatezza delle scelte effettuate senza tenere conto di altri fattori, in particolare della distribuzione dei prodotti ai revisori e della stabilità e dell'omogeneità dei loro giudizi.

5.1.2 La distribuzione dei prodotti ai revisori

Nell'Area delle Scienze Politiche e Sociali l'assegnazione dei prodotti ai revisori è stata effettuata rispettando quanto previsto nel documento relativo ai criteri: «ove possibile, l'individuazione dei 2 revisori peer verrà fatta separatamente da 2 membri distinti del GEV di riferimento» (Anvur, 2013d, GEV14, Appendici, p. 73). Come già esposto (*cfr*. Capitolo 1) cioè, ciascun prodotto da valutare veniva assegnato a due diversi esperti valutatori (EV) all'interno del GEV che,

[165] Alcune di queste questioni sono sollevate anche nelle interviste: «il problema è che, ripeto, pochi valutatori, pochissimi su alcuni argomenti, pochissimi, e molto spesso in conflitto. Perché quando sono pochissimi, sono quelli, e ci sono delle discipline e quelli sono. Questo ha creato conflitti di interessi, oppure conflittualità interna. Nel senso che indipendentemente dal conflitto di interesse ci poteva essere, si è verificato spesso, come in tutte le *peer review* una, diciamo così, antipatia verso il personaggio, oppure un'eccessiva simpatia, e quindi una valutazione squilibrata. Questo però non c'è modo di evitarlo nella misura in cui si passa per la peer review, perché… l'unica cosa è che se era proprio discordante poi alla fine c'era questo sistema di compensazione, di aggiusto. Questo sicuramente è un problema che c'è stato. Piccole guerre accademiche, magari, non esagerate ecco. Direi che mi sarei aspettata di peggio, non so se per le mie discipline, ma non c'è stata questa… per alcune discipline c'è stata una vera e propria guerra accademica. Nell'Area 14, che era la nostra, io faccio parte dei politologi, c'è stata una certa durezza da parte di gruppi politologi che sono dominanti rispetto ad altre aree della… disciplina… aree anche geopolitiche, aree territoriali che sono state penalizzate. Ma questo è avvenuto per l'altro gruppo, la sociologia ancora di più, cioè è stata molto rigida, perché non c'è un modo unico, insomma i criteri sono sempre quelli: da valutare l'originalità, l'internazionalizzazione… però è un po' difficile la valutazione che non sia oggettiva, legata a degli indicatori stabili, non c'erano indicatori stabili» (Intervista Bazzicalupo).

indipendentemente l'uno dall'altro, selezionavano un revisore dall'albo, oppure ne proponevano uno qualora nell'albo non fossero presenti profili adeguati[166].

Il criterio fondamentale, pur non esplicitato nei rapporti finali, era il settore scientifico-disciplinare di riferimento sia per l'assegnazione dei prodotti agli EV, sia per l'assegnazione da parte di questi dei prodotti ai revisori[167].

I criteri prevedevano, inoltre, una serie di principi di cui il GEV doveva tenere conto nella selezione dei revisori: «la scelta dei revisori esterni verrà effettuata evitando conflitti di interesse tra i revisori stessi e gli autori e/o la struttura di affiliazione. Inoltre, verrà garantita l'indipendenza dei revisori ponendo attenzione alla sede di affiliazione, alla collaborazione scientifica, e, ove possibile, alla nazionalità. Per minimizzare i conflitti di interesse, si privilegeranno i revisori operanti al di fuori dei confini nazionali» (*ibidem*). Nel rapporto finale si sottolinea il rispetto di questi criteri di assegnazione e in una nota si sottolinea che «in alcuni casi in cui l'assegnazione non ha tenuto conto dei criteri suddetti, gli stessi *referee* hanno segnalato il problema permettendo di ovviare all'errore» (Anvur, 2013d, GEV14, p. 19 nota 4). Purtroppo non vi sono dati né sul numero di casi in cui l'assegnazione non ha tenuto conto dei criteri previsti, né sul numero di assegnazioni riviste a seguito delle segnalazioni dei revisori[168].

Osservando i numeri relativi alle revisioni effettuate per SSD del prodotto (Tabella 40) risulta evidente lo squilibrio della distribuzione: evidentemente i settori disciplinari con più prodotti hanno richiesto un numero maggiore di revisioni, ma non sono disponibili dati sul numero di revisori impiegati per ogni SSD. Chiaramente un revisore potrebbe aver valutato più prodotti in uno stesso settore disciplinare, oppure in settori disciplinari differenti, ciò nonostante una tabella descrittiva sarebbe risultata estremamente utile per la rendicontazione delle risorse impiegate in ciascun settore.

---

[166] Una procedura mirata al bilanciamento di eventuali distorsioni: «nessun lavoro è stato assegnato interamente ad un solo componente del GEV, ma ogni componente del GEV poteva fare solo una assegnazione. L'altra assegnazione, all'altro *referee*, per la seconda valutazione avveniva all'insaputa della prima da un altro componente del GEV, garantendo in questo modo, quantomeno, che non ci fosse quel bias dovuto per esempio all'antipatia-simpatia, odio-amore, ostilità nei confronti della scuola piuttosto che alleanza nei confronti della scuola. Questo noi pensiamo di averlo abbastanza evitato, ed è stato un secondo criterio di qualità che abbiamo utilizzato» (Intervista Colozzi).

[167] Ai fini dell'assegnazione: «abbiamo guardato gli SPS presenti nel GEV, abbiamo suddiviso i prodotti. Ci potevano essere, per esempio, due 08, due 07, due 010 due 04, allora su quella base abbiamo distribuito i prodotti. Quelli di 04 si sono distribuiti i prodotti di 04, quelli di 08 i prodotti di 08, eccetera» (Intervista Colozzi).

[168] Le informazioni disponibili per gli EV al momento dell'assegnazione erano diverse: "[i nomi dei revisori] sono poi stati inseriti in un database del CINECA su cui si poteva fare una ricerca per settore scientifico-disciplinare, oppure per area scientifica ERC, a quel punto venivano proposti dei nomi ai membri GEV con anche l'indicazione di eventuali conflitti di interesse, ad esempio il fatto che fosse della stessa università o che ci fosse stato un coautoraggio visibile nel sito del CINECA, cosa non tanto facile per la verità. Se poi questi nomi non soddisfacevano il membro GEV lui poteva aggiungerne degli altri. Veniva anche indicato quante revisioni erano già state affidate a quella persona, non solo, ma ciascun revisore aveva indicato un limite superiore al numero di revisioni che poteva effettuare. Queste erano le informazioni che venivano utilizzate […] quando il revisore accettava di fare il revisore GEV indicava delle sue aree. Quindi c'erano le aree legate al prodotto da rivedere, il settore scientifico disciplinare, poi c'era il settore ERC, sempre legato al prodotto, e poi c'era la selezione dei nomi da proporre al membro GEV, che veniva proprio fatta correlando le informazioni del prodotto con quelle che il revisore aveva indicato» (Intervista Benedetto).

Tabella 40 - Revisioni per SSD (revisori ripetuti in ogni SSD di competenza) per nazionalità di affiliazione (adattamento della tabella 2.11, Anvur, 2013d, GEV14, p. 28)

SSD	Revisioni effettuate da revisori con affiliazione italiana			Revisioni effettuate da revisori con affiliazione straniera		
	n	% riga	% colonna	n	% riga	% colonna
SPS/01	474	84,5	6,9	87	15,5	5,3
SPS/02	551	91,1	8,0	54	8,9	3,3
SPS/03	276	96,5	4,0	10	3,5	0,6
SPS/04	700	71,2	10,2	283	28,8	17,1
SPS/05	117	99,2	1,7	1	0,8	0,1
SPS/06	292	93,3	4,3	21	6,7	1,3
SPS/07	1.621	81,3	23,6	372	18,7	22,5
SPS/08	1.263	80,0	18,4	315	20,0	19,0
SPS/09	580	70,8	8,5	239	29,2	14,4
SPS/10	289	74,3	4,2	100	25,7	6,0
SPS/11	179	64,6	2,6	98	35,4	5,9
SPS/12	278	79,9	4,1	70	20,1	4,2
SPS/13	145	100	2,1			
SPS/14	93	95,9	1,4	4	4,1	0,2
Totale	**6.858**	**80,6**	**100,0**	**1.654**	**19,4**	**100,0**

La variabilità del numero di revisioni effettuate da ciascun revisore (Tabella 41) risulta elevatissima: la distribuzione del numero di revisori italiani per classe è piuttosto eterogenea (l'indice di Gini risulta pari a 0,83), leggermente meno quella dei revisori stranieri (0,79)[169]. Non essendo stato fissato un numero minimo o un numero massimo di revisioni né programmato il numero di revisori da coinvolgere per ciascun settore (senza contare la situazione sostanzialmente emergenziale in cui le assegnazioni sono state effettuate nelle ultime fasi dell'esercizio di valutazione), il numero di revisioni effettuate da ciascun revisore è risultato estremamente disomogeneo.

Tabella 41 - Revisioni effettuate in classi per affiliazione del revisore (Tab. 2.12, Anvur, 2013d, GEV14, p. 29)

Revisioni in classi	*Revisori con affiliazione italiana*		*Revisori con affiliazione straniera*		*Totale*	
	n	*%*	*n*	*%*	*n*	*%*
1-5 revisioni	57	*14,00*	41	*27,15*	98	*17,56*
6-10 revisioni	64	*15,72*	39	*25,83*	103	*18,46*
11-15 revisioni	83	*20,39*	33	*21,85*	116	*20,79*
16-20 revisioni	67	*16,46*	21	*13,91*	88	*15,77*
21-25 revisioni	48	*11,79*	10	*6,62*	58	*10,39*
>25 revisioni	88	*21,62*	7	*4,64*	95	*17,03*
Totale	**407**	***100,00***	**151**	***100,00***	**558**	***100,00***

Uno studio sulla valutazione peer delle proposte di finanziamento ARC (*Australian Research Council*) ha rilevato che i giudizi risultano migliori quando i revisori hanno un piccolo numero di elementi da valutare (Jayasinghe *et al.* 2003). Questo effetto potrebbe derivare dalla mancata possibilità di confrontare i prodotti da valutare con un set più ampio, dunque di effettuare una sorta di *benchmarking* nel corso della valutazione, anche con specifico riferimento a uno o più campi disciplinari. Con riferimento alla VQR, data l'elevata variabilità del numero di revisioni per revisore e del numero di revisioni in base al SSD di riferimento, sarebbe stata dunque opportuna la

[169] Grazie all'intervista al professor Benedetto è possibile affermare che i *referee* avevano facoltà di segnalare il numero massimo di prodotti che era disposto a revisionare (cfr. nota 168), tuttavia a livello centrale o di GEV non era stato fissato alcun limite minimo o massimo alle revisioni effettuabili da ciascun revisore.

pubblicazione di dati più chiari e dettagliati. L'unica spiegazione riportata fa riferimento alle diverse disponibilità dei revisori, oltre che alla constatazione che alcuni di essi si sono fatti carico anche di prodotti rifiutati o valutazioni inevase, tuttavia le conseguenze di questa variabilità non vengono esposte né in forma di ipotesi né, tantomeno, a seguito di analisi mirate[170].

Vi sono altre possibili distorsioni connesse all'attribuzione dei prodotti ai revisori e dovute alle caratteristiche di revisori, autori e prodotti in particolare il campo di studi, il prestigio e il ruolo accademico (Crane, 1967; Zuckerman e Merton, 1973; Mahoney, 1977; Benos *et al.* 2007). A proposito del campo di studi sono già state evidenziate diverse peculiarità delle scienze politiche e sociali, in grado di introdurre distorsioni di vario genere nella valutazione dei prodotti da parte di un revisore (*cfr.* § 5.2.1)[171]. Il prestigio e il ruolo accademico dei revisori andrebbero considerati con altrettanta attenzione. Il dibattito circa le influenze di prestigio e ruolo nel riconoscimento dei meriti scientifici è di origini mertoniane (Merton, 1968) ed evidenzia rischi differenti a seconda della direzione dell'asimmetria:

a) in una situazione in cui il prestigio o il ruolo accademico dell'autore sono molto elevati rispetto a quelli del revisore quest'ultimo potrebbe (più o meno inconsciamente) spingere in alto la propria valutazione, anche non in relazione con il prodotto da valutare ma per acquiescenza o riconoscimento di meriti più generali;
b) in una situazione in cui il prestigio o il ruolo accademico del revisore sono molto elevati rispetto a quelli dell'autore, invece, il revisore potrebbe affrontare la revisione senza grosse aspettative e (più o meno inconsciamente) spingere verso il basso la propria valutazione.

Con riferimento alle disuguaglianze nel mondo scientifico è più che nota la individuazione da parte di Merton del cosìdetto effetto San Matteo, per cui «a chi ha, sarà dato e sarà nell'abbondanza, e a chi non ha sarà tolto anche quello che ha» (Vangelo di Matteo, 13:12). Questo effetto è

[170] L'opinione espressa spontaneamente da uno dei membri del GEV 3 su questa questione è estremamente chiara: «il problema, l'unico problema che ho intravisto è che mentre un panel, e questo è vero in tutti i processi di valutazione, vede tutta la produzione contemporaneamente e quindi si fa un quadro di quello che è il valor medio e può capire che cos'è sopra la media e che cos'è sotto la media; il singolo revisore che vede un singolo prodotto fa una valutazione che è individuale ma senza nessun parametro di raffronto, questo può sparare un prodotto verso l'alto o verso il basso indipendentemente dalla buona o cattiva volontà del revisore, è che non ha nessun punto di riferimento. È vero che molti revisori hanno ricevuto diversi prodotti da valutare quindi potevano farsi un loro ranking interno, alcuni ne hanno avuti pochi eccetera. Quindi questa parte è ovviamente la parte più delicata, cosa che ovviamente penso che in settori non bibliometrici abbiano sofferto ancora di più» (Intervista Pacchioni).

[171] Diversi stralci di intervista riguardano questo aspetto, sia con riferimento all'intero esercizio: «noi abbiamo utilizzato tutta una serie di garanzie per evitare i conflitti di interesse, però le garanzie riguardavano fattori oggettivi: la stessa università, hanno collaborato in passato, eccetera, certamente non andavano invece a individuare altre cose come essere della stessa scuola oppure non esserlo, avere avuto problemi o amicizie passate. Insomma la peer review è un problema soprattutto quando coinvolge tantissimi prodotti da rivedere, ha sicuramente un problema di uniformità di giudizio…» (Intervista Benedetto); che con riferimento all'Area 14: «sicuramente sono state messe in atto delle procedure per far si che questi processi quali ad esempio la scelta dei revisori, giusto per citarne uno, svolgessero secondo i criteri che fossero scelti consapevolmente dal GEV e quindi per esempio c'è stata un'attenzione in tutti i GEV e anche nel nostro al problema del conflitto di interessi o comunque anche questa cosa che diceva il direttore [*ndr. Torrini*] di diversificare l'affiliazione dei revisori per far si che tutte le componenti scientifiche, politiche e accademiche di qualsiasi natura fossero in qualche modo rappresentate nel processo di peer review e che il processo cioè questo pluralismo fosse la garanzia principale della qualità poi dell'output della valutazione» (Intervista Blasi).

conosciuto soprattutto in relazione all'analisi citazionale, ma altrettanto riscontrabile in relazione all'attribuzione di riconoscimenti e premi ed alla stessa possibilità di pubblicare (Merton 1968; 1968a; 1988; Crane, 1965). Merton, inoltre, non ha mancato di sottolineare come siano riscontrabili distribuzioni disuguali di produttività anche in riferimento alle istituzioni scientifiche (1968), un effetto in seguito definito come vantaggio cumulativo istituzionale (Bentley e Blackburn, 1990). Vale la pena a questo proposito ricordare anche l'effetto alone (*halo effect*), per cui, ad esempio, scienziati che lavorano presso università o istituti particolarmente prestigiosi hanno una maggiore produttività, perché i loro articoli vengono accettati più facilmente dalle riviste scientifiche (Crane, 1965). Dunque prestigio e ruolo non sono le uniche caratteristiche da tenere in considerazione, la stessa struttura di afferenza degli autori potrebbe dar luogo a un vantaggio o uno svantaggio nella valutazione a seconda anche delle caratteristiche del revisore (in particolare della sua percezione del prestigio dell'istituzione). E' inoltre nota l'influenza del nepotismo sugli esiti delle procedure di peer review (si vedano ad esempio Wennerâs e Wold, 1997; Sandström e Hällsten, 2008). Le accortezze adottate dal GEV in riferimento ai conflitti di interesse e quelle messe in atto nell'assegnazione dei prodotti dovrebbero avere almeno in parte ridotto il rischio che le valutazioni dei prodotti nel corso della VQR risentissero di questo genere di distorsione[172].

Il dibattito circa le distorsioni legate al genere è invece sostanzialmente aperto; mentre infatti alcuni studi hanno riscontrato differenze significative nell'esito della peer review in base al genere di autori e revisori (Lloyd, 1990; Wennerâs e Wold, 1997; Sandström e Hällsten, 2008), altri non hanno evidenziato differenze (Gilbert *et al.* 1994; Grant *et al.* 1997; si veda anche Benos *et al.* 2007). Bormann e colleghi (2007), a seguito di una meta-analisi riferita a 21 studi, sostengono l'evidenza di robuste differenze di genere nelle procedure di selezione per i finanziamenti (nel modello stimato la probabilità di ricevere fondi per gli uomini è del 7% superiore rispetto alla probabilità per le donne), mentre Ceci e Williams (2011) suggeriscono che la sottorappresentazione delle donne nelle comunità scientifiche e più in particolare nel panorama delle pubblicazioni sia dovuta più a una differenza di accesso alle risorse che a una distorsione di genere nella peer review.

Molte delle possibili distorsioni appena esposte sono state analizzate con riferimento alla valutazione dei prodotti nella VTR (Reale *et al.* 2007), condotta dal CIVR interamente tramite peer review (*cfr.* Capitolo 1). La VTR è molto interessante ai fini di questo studio, sotto diversi punti di vista infatti risulta estremamente simile alla VQR: si tratta un esercizio di valutazione istituzionale, svolto ex-post, cioè a seguito della pubblicazione, e solo su una parte della produzione scientifica; inoltre è basato sugli stessi criteri ed essenzialmente mirato agli stessi scopi, per quanto le sue dimensioni risultino decisamente ridotte rispetto a quelle della VQR (*cfr*. § 1.2 e 1.3).

L'analisi presentata dalla Reale e dal suo gruppo mira a valutare diversi aspetti della peer review: razionalità, affidabilità, imparzialità, efficienza, efficacia, controllando la presenza, ed eventualmente l'entità, di alcune possibili distorsioni. I giudizi dei pari forniti nel corso della VTR sono analizzati allo scopo di individuare le distorsioni legate al prestigio delle istituzioni, alla reputazione dei ricercatori, alla ricerca interdisciplinare, e all'accordo tra i revisori, con riferimento a quattro macro-aree: chimica, biologia, scienze umane ed economia (Reale *et al.* 2007). Il prestigio delle

[172] Emerge sia dall'intervista al coordinatore che dalle interviste ai membri del GEV 14 un particolare attenzione a questi aspetti, si veda ad esempio la nota 168 a p. 134.

istituzioni e la reputazione dei ricercatori[173] vengono presentati come ininfluenti rispetto ai giudizi dei pari: rispetto al prestigio delle istituzioni si riscontrano campi di variazione dei giudizi ampi, mentre la reputazione dei ricercatori non risulta significativamente associata con i giudizi (Reale *et al.* 2007). La questione della ricerca interdisciplinare viene analizzata confrontando i risultati ottenuti dalle università con quelli ottenuti dagli enti di ricerca, generalmente più inclini alla ricerca interdisciplinare (Reale *et al.* 2007). L'analisi mostra una differenza significativa tra i risultati di questi due tipi di istituti, a sfavore degli enti di ricerca, evidenza che secondo gli autori suggerisce la presenza di un bias nei confronti della ricerca interdisciplinare.

L'analisi proposta è evidentemente limitata dalla quantità e dal genere di dati disponibili, gli autori stessi evidenziano come lo studio non indaghi aspetti come la discriminazione, le distorsioni legate ai diversi approcci teorici e metodologici, né la stabilità nel tempo delle valutazioni. Sarebbe auspicabile effettuare simili e più approfonditi controlli sui giudizi espressi dai revisori nel corso della VQR. Il numero di prodotti sottoposti a valutazione in questo secondo esercizio e la minore discrezionalità degli enti nella selezione dei prodotti potrebbe infatti dar luogo a risultati estremamente diversi.

Inevitabilmente, in relazione a questi e altri fattori, vanno considerate le possibili conseguenze di una valutazione che non preveda l'oscuramento né degli autori dei prodotti né della loro collocazione editoriale. Attualmente la forma di peer review più diffusa, adottata anche in occasione della VQR 2004-2010, prevede che i revisori non siano all'oscuro delle identità degli autori o delle loro affiliazioni, mentre le identità dei revisori stessi non vengono rivelate né agli autori né ad altri revisori. Si tratta di una *blind review*, mentre per ottenere una *double blind review* sarebbe necessario l'oscuramento delle identità e delle affiliazioni degli autori; sul versante opposto una *open review* prevedrebbe lo svelamento le identità dei revisori, e una completa trasparenza e accessibilità per l'intera procedura.

Lo scopo dell'oscuramento degli autori è naturalmente la riduzione dei rischi di distorsione nel processo di valutazione: non conoscendo autore e affiliazione il revisore non dovrebbe essere influenzato da precognizioni o pregiudizi nella sua valutazione. Nella pratica è però estremamente difficile oscurare efficacemente autori e affiliazioni (Benos *et al.* 2007). Si pensi ad esempio a quanto la bibliografia possa tradire l'identità dell'autore, se non personale istituzionale, e a quanto questo stesso elemento sia fondamentale alla valutazione della qualità di un prodotto. Se ciò è vero prima della pubblicazione lo è a maggior ragione dopo la pubblicazione stessa, quando i revisori potrebbero non solo reperire facilmente le informazioni oscurate, ma addirittura conoscerle già, avendo già letto il lavoro. Nel corso di un esercizio di valutazione come la VQR l'oscuramento degli autori diventa estremamente difficoltoso, se non impossibile. Richiederebbe infatti non solo l'oscuramento di autori e affiliazioni, ma anche quello della sede editoriale e del titolo del lavoro, tutti identificativi in grado di condurre facilmente i revisori alle identità degli autori, sempre che non ne siano già a conoscenza.

Circa questo punto, uno studio di Cho e colleghi (1998) ha segnalato che nel caso vengano oscurati solo i nominativi degli autori e le loro affiliazioni l'oscuramento risulta efficace in circa il 60% dei casi. I revisori che riuscivano ad identificare gli autori a dispetto dell'oscuramento risultavano in

[173] Il prestigio delle istituzioni è operativizzato in termini di dimensioni ed età delle istituzioni, controllata con il confronto con il ranking ARWU (Academic Ranking of World Universities, stilato annualmente dall'Università Jiao Tong di Shanghai); la reputazione dei ricercatori è operativizzata facendo riferimento alla posizione accademica (Reale et al., 2007).

generale più produttivi e con una maggiore esperienza di ricerca. In altri termini si potrebbe dire che tanto più un revisore è partecipe del dibattito scientifico e della produzione scientifica nella propria disciplina tanto più è probabile che riesca ad identificare l'autore dei lavori che gli vengono sottoposti, anche in una *double blind review*.

Gli studi effettuati sull'oscuramento, in campo editoriale, hanno dato risultati contrastanti. McNutt e colleghi (1990) hanno chiesto agli editori di valutare la qualità delle revisioni, che è risultata più alta nel caso di revisioni *blinded*[174]. Il risultato potrebbe tuttavia dipendere dal fatto che i revisori stessi erano a conoscenza della loro partecipazione a uno studio[175].

In riferimento al ruolo e al prestigio, anche istituzionale, Peters e Ceci (1982), una volta oscurati gli autori, hanno individuato quote elevate di rifiuti per articoli pubblicati da stimati gruppi di ricerca, ma uno studio più recente (Garfunkel *et al.* 1994) ha indicato una mancanza di associazione tra lo status istituzionale e l'accettazione di articoli, presente invece in relazione a brevi rapporti di ricerca. Justice e colleghi (1998) hanno riscontrato una maggiore difficoltà a ottenere un oscuramento efficace per gli autori più conosciuti, senza peraltro individuare una relazione tra la qualità della revisione e l'oscuramento dell'autore[176]. L'ipotesi avanzata è semplice: le revisioni che dovrebbero risentire maggiormente dell'oscuramento sono quelle provenienti da autori prestigiosi, dunque il fatto che proprio questo genere di autore risulti più semplice da identificare può avere inciso sulla mancanza di effetto dell'oscuramento[177], inoltre il numero delle identificazioni degli autori potrebbe essere maggiore nei campi di studio più specialistici (Lane, 2008).

Una revisione aperta in cui neppure l'identità dei revisori venga oscurata potrebbe ridurre o accentuare ulteriormente i rischi di distorsione connessi al ruolo e al prestigio. Da un lato, ridurrebbe le possibilità di nascondere conflitti di interesse e renderebbe più trasparente il processo, spingendo verosimilmente i revisori a dedicare più attenzione alla valutazione del prodotto[178]. Dall'altro, però, l'oscuramento dei revisori ha una funzione di protezione che verrebbe a mancare: una revisione aperta potrebbe dar luogo a revisioni meno critiche sia da parte di revisori esperti e prestigiosi (che potrebbero temere di offendere i propri pari) sia nel caso di revisori giovani (che potrebbero temere di risultare eccessivamente presuntuosi o ostili).

Non mancano gli studi in questo campo. Van Rooyen e il suo gruppo (1998), ad esempio, hanno individuato giudizi più favorevoli nel caso di revisioni aperte, rispetto alle revisioni effettuate da *referee* anonimi. La differenza sembra poco significativa, tuttavia i risultati potrebbero variare a

[174] In questo studio l'oscuramento delle revisioni è stato efficace nel 73% dei casi. Lo strumento utilizzato per la rilevazione della qualità delle revisioni era una scala likert, a cinque gradienti, simile a quello successivamente messo a punto da van Rooyen e il suo gruppo (1999; *cfr*. nota 179, p. 140).

[175] Sono noti diversi effetti di interazione che agiscono soprattutto sui soggetti delle indagini, si pensi all'effetto Hawthorne, al *guinea-pig effect* o all'effetto placebo (Mayo, 1933; Merton, 1968; Campbell e Stanley; 1966; Campbell e Cook, 1979), ma vi sono anche meccanismi che coinvolgono sia i soggetti che i ricercatori, come l'effetto Rosenthal e il teorema di Thomas (Rosenthal e Jacobson, 1968; Merton, 1968). Si noti che in una più ampia analisi di van Rooyen e colleghi (1998; 1999) non sono emerse significative differenze in relazione alla qualità delle revisioni, tuttavia l'oscuramento è risultato efficace solo nel 51,2% dei casi.

[176] Lo studio, sperimentale, evidenziava un *odd ratio* di 0,3 per il successo nell'oscuramento degli autori più noti rispetto agli altri.

[177] Risultati simili sono stati ottenuti in altri studi (Ceci e Peters, 1984; Yankauer, 1991; Baggs et al., 2008).

[178] E' l'opinione ad esempio della professoressa Bazzicalupo: «la responsabilità dei valutatori dovrebbe essere un po' più chiara, perché questa non c'è. Purtroppo non ci sta per via del fatto che il meccanismo è occulto e tutti sono… però un sistema di responsabilizzazione ci dovrebbe essere, magari da parte di colui che fa l'attribuzione e risulterà responsabile dei giudizi» (Intervista Bazzicalupo).

seconda dell'ampiezza del campo di studi: tanto più specialistico è il campo tanto più è probabile che revisori e autori si conoscano personalmente. In un'altra analisi dello stesso gruppo (van Rooyen *et al.* 1999) non si sono evidenziate differenze nella qualità delle revisioni[179], tuttavia l'esito delle revisioni risultava negativo con il 12% di probabilità nel caso di una procedura aperta[180]. McNutt e colleghi (1990) hanno invece osservato revisioni più cortesi e costruttive nonché una quota più alta di revisioni positive da parte dei revisori che accettano di firmare.

In sintesi, una revisione aperta non influisce né sull'esito né sulla qualità delle revisioni, tuttavia i revisori sono meno disposti ad accettare revisioni in questa modalità, e la loro disponibilità a firmare le revisioni è più alta nel caso in cui queste risultano positive. La letteratura dunque non evidenzia effetti positivi legati a una revisione aperta, che sembrerebbe invece portare a un aumento dei rifiuti delle revisioni e dunque dei conseguenti rischi di distorsione.

A tutti i possibili fattori in grado di distorcere i giudizi si aggiunge una questione essenziale: la capacità valutativa. A questo proposito Davidoff (1998), ha rilevato come le problematiche evidenziate dagli studi sulla peer review sembrino connesse più alla mancanza di capacità di valutazione critica che a distorsioni sistematiche, opinioni inappropriate o ipercriticismo dei revisori[181].

Baxt e colleghi (1998) hanno effettuato uno studio *ad hoc* mirato a valutare la performance dei revisori utilizzando come strumento un manoscritto appositamente modificato con l'introduzione di errori, e ben il 68% dei revisori non ha notato che i risultati addotti non supportavano le conclusioni tratte nel manoscritto[182].

Il nodo centrale per l'affidabilità della peer review dunque potrebbe essere identificato nell'esperienza dei revisori o in alternativa nella loro preparazione al compito. Uno studio sperimentale condotto sulla preparazione dei revisori (Schroter *et al.* 2004) ha effettivamente

[179] Lo strumento utilizzato per valutare la qualità delle revisioni è l'RQI (R*eview Quality Instument*), che mira a rilevare con una scala Likert da sottoporre agli editori quanto un revisore ha tenuto conto di cinque aspetti del prodotto da valutare (l'importanza del quesito di ricerca, l'originalità, i punti di forza e di debolezza del metodo, la presentazione e l'interpretazione dei risultati) e a valutare due aspetti della revisione stessa (la costruttività e la fondatezza dei commenti).

[180] Neppure Godlee e il suo gruppo (1999) hanno riscontrato differenze significative nella qualità delle revisioni o nella capacità di identificare gli errori nel caso ai revisori venisse chiesto di firmare la revisione, tuttavia solo la metà dei revisori accettò di firmare.

[181] Vanno in questa direzione alcune delle osservazioni avanzate dal professor Colozzi: «una delle ipotesi interpretative che io ho fatto […] è che da noi la pratica del referaggio è ancora molto poco diffusa. Se lei pensa, il referaggio delle riviste è diventato abituale negli ultimi tre-quattro anni, prima praticamente non esisteva se non in pochissime riviste, alcune delle quali lo dichiaravano e non lo facevano, tra l'altro, diciamo che in ogni caso era molto limitato. E' escluso ancora quasi totalmente dalle monografie. Quindi in realtà gli italiani che hanno fatto un lavoro di referaggio serio non sono poi così tanti, e soprattutto quelli che lo hanno fatto non hanno esperienza europea, non hanno esperienza internazionale, perché quei pochi che lo hanno fatto è perché facevano già da *referee* delle riviste ad impact factor che si rivolgevano a loro per prodotti magari di italiani o comunque su tematiche… venivano coinvolti dalle direzioni di queste riviste. Tranne questa piccolissima cerchia di persone che ha maturato una esperienza internazionale e quindi ha un'idea di standard di qualità, per gli italiani questo standard probabilmente non c'è» (Intervista Colozzi).

[182] Il manoscritto è stato revisionato da 203 revisori di *Annals of Emergency Medicine* (il 78% del totale), rifiutato da 117 (che hanno individuato circa il 39% degli errori principali e il 25% di quelli secondari), accettato da 15 (che hanno individuato solo il 17% degli errori principali e l'11% di quelli secondari) e raccomandato per ulteriori revisioni da 67 (che hanno individuato circa il 29% degli errori principali e il 22% di quelli secondari). Evidentemente la raccomandazione per l'accettazione, il rifiuto o l'ulteriore revisione dipende dal numero di errori individuati.

mostrato una maggiore qualità delle revisioni e una maggiore capacità di individuare gli errori da parte dei revisori che erano stati sottoposti ad un breve addestramento[183]. Nonostante l'impatto non fosse di grande entità[184] né di lunga durata (dopo soli sei mesi le differenze tra i gruppi non erano più riscontrabili), è chiaro che una migliore preparazione dei revisori è l'unico mezzo efficace per ottenere revisioni più affidabili.

5.1.3 Omogeneità e stabilità dei giudizi e delle scale di valutazione

La selezione dei revisori e l'assegnazione dei prodotti da valutare sono momenti fondamentali del processo di valutazione. Un'altra questione, meno semplice sia da analizzare che da valutare, è relativa alle scale di giudizio utilizzate dai revisori per la valutazione dei prodotti.

La pratica della peer review si basa sul presupposto essenziale che i revisori, in quanto appartenenti a una stessa comunità scientifica, condividano una serie di conoscenze e valori su cui basare la valutazione e che dunque i loro pareri si fondino su un'unica scala di giudizio. Le scale di giudizio dei revisori dovrebbero dunque risultare, in teoria, omogenee (diversi revisori dovrebbero condividere la stessa scala di giudizio) e stabili (ciascun revisore in momenti diversi dovrebbe fare riferimento alla stessa scala di giudizio). Ciò dovrebbe tradursi in valori elevati di accordo per i giudizi (*intra* ed *inter rater agreement,* Hojat *et al.* 2003), eppure una delle critiche più diffuse alla pratica della peer review fa riferimento proprio alla mancanza di un livello di accordo accettabile tra i revisori (Marsh e Ball, 1981; Cicchetti, 1991; Callaham *et al.* 1998; Jayasinghe, 2003; Hojat *et al.* 2003).

A questo proposito è interessante notare la similarità di questi presupposti con le forme che l'attendibilità può assumere con riferimento alla tecnica dell'analisi del contenuto: intra-rilevatore ed inter-rilevatore, cui si aggiunge la convergenza con uno standard conosciuto e valido (Krippendorff, 1980; Nobile, 1997). E' proprio a causa della difficoltà di disporre di uno standard validato, oltre che per le problematiche logiche, matematiche o semantiche implicate dall'uso di altri criteri, che «l'attenzione dei metodologi riservata alle tecniche di controllo dell'attendibilità dell'analisi del contenuto è stata fin ora piuttosto avara e comunque in gran parte incentrata sull'elaborazione di coefficienti di concordanza che dessero conto dell'attendibilità intra ed inter codificatore» (Nobile, 1997, p. 103).

March e Ball (1981) hanno definito l'attendibilità dei revisori (*single-rater reliability*) come la correlazione tra i punteggi assegnati da due revisori indipendenti a uno stesso elemento (su un ampio numero di elementi). Hojat e colleghi (2003) hanno brevemente passato in rassegna gli studi sull'accordo dei revisori, segnalando sia studi che fanno riferimento a un modesto grado di accordo, intorno al 20% (Hendrick, 1976; Scott, 1974), sia studi con un livello di accordo più elevato, attorno al 50% (Crandall, 1978; Scarr and Weber, 1978; Cicchetti, 1980)[185]. A questo proposito Bornmann, Mutz

[183] Lo strumento proposto agli editori per valutare le revisioni è nuovamente l'RQI (*Review Quality Instument*), proposto da van Rooyen et al. (1999), cfr. nota 179, p. 140.

[184] Il punteggio medio per i gruppi sperimentali (uno istruito attraverso un'auto-formazione l'altro faccia a faccia) pari a 2,85 e 2,75 contro il 2,56 del gruppo di controllo su una scala da 1 a 5.

[185] In riferimento alle riviste, sulla base di una rassegna di ricerche precedenti, March e Ball riportano una *single-rater reliability* di 0,27. Sempre basandosi anche su studi precedenti in una sorta di meta analisi, con riferimento alle riviste Cicchetti (1991) riporta valori tra lo 0,19 e lo 0, 54 (con una mediana di 0,30) mentre,

e Daniel (2010), a seguito di una meta-analisi su 48 studi, hanno evidenziato che le indagini che mostrano alti livelli di accordo sono meno affidabili statisticamente di quelli che sottolineano un basso livello di accordo.

Del resto un completo accordo nella peer review non è auspicabile; se, da un lato, rassicurerebbe circa l'univocità delle scale di giudizio, dall'altro, creerebbe un rischio di ridondanza delle valutazioni (Cole *et al.* 1981; Balair, 1991; Langfeldt, 2006; Bornmann, 2008). Inoltre è chiaro il legame tra la mancanza di accordo e l'appartenenza a differenti posizioni teoriche o metodologiche (paradigmi), e aree di competenza tra i revisori (Chubin e Hackett, 1990; Eckberg, 1991; Kostoff, 1995; Laudel; 2006). Questi ultimi due argomenti rivestono un ruolo centrale con riferimento alle scienze sociali proprio in relazione alle loro peculiarità in termini di relazioni tra specializzazioni, approcci metodologici e teorici, di interdisciplinarietà, di collaborazione scientifica e di consenso (Moody, 2004).

Jayasinghe e colleghi (2003), sulla base delle osservazioni di Lindsey (1978) circa le peculiarità delle scienze sociali e le loro possibili conseguenze sull'esito della peer review, hanno analizzato separatamente i risultati delle revisioni nell'area delle scienze sociali, senza però riscontrare un grado di accordo inferiore a quello rilevato per le scienze dure. I livelli di accordo riscontrati in letteratura non sono dunque elevatissimi, ma almeno su questo piano sembrano esserci rassicurazioni circa le eventuali conseguenze derivanti dalle peculiarità delle scienze sociali.

Nel caso della VQR non è stato pubblicato alcun dato complessivo sul grado di accordo tra i revisori, né a latere del rapporto finale né a latere dei rapporti di Area[186]. Le dimensioni dell'esercizio di valutazione e le procedure utilizzate complicano enormemente questo tipo di analisi (si pensi alla varietà dei tipi di prodotti, alle variazioni nel numero di revisioni effettuate, al numero di settori scientifico-disciplinari di riferimento per prodotti e revisori), ma la rendono se possibile ancora più interessante. Non si tratta esclusivamente di rendicontare le procedure di valutazione alla comunità scientifica, ma si tratta anche di restituire elementi di riflessione sulla condivisione dei criteri valutativi, sul consenso scientifico, sul grado di coesione della comunità che possano rivelarsi utili alla riflessione sulle pratiche, sul dibattito e sulle modalità della comunicazione scientifica[187].

---

circa le proposte per il finanziamento, lo stesso Cicchetti (*ibidem*) riporta valori di *single-rater reliability* tra lo 0,17 e lo 0,37 (con una mediana di 0,33).

[186] Sembrerebbe, da quanto rilevato nel corso delle interviste, che una volta conclusa la procedura siano stati effettuati diversi controlli sui giudizi espressi dai revisori, ad esempio il Presidente Fantoni afferma che: «abbiamo anche poi, in una fase successiva, cercato di analizzare il comportamento dei *referee*» (Intervista Fantoni) e il professor Bonaccorsi fa riferimento a dei controlli ex-post; tuttavia nessun dato a riguardo è ancora stato pubblicato.

[187] E' interessante la riflessione del professor Bonaccorsi sulle valutazioni ottenute dalle monografie in alcune Aree: «i punteggi ci suggeriscono che ci sia stata una vera e propria crisi della monografia, ma non dovuta a una distorsione della valutazione, dovuta a una dinamica interna alle discipline che hanno fatto troppo a lungo sopravvivere la monografia come uno strumento essenziale per i passaggi concorsuali senza riflettere sul loro contenuto di originalità. In discipline che sempre più sono internazionalizzate, hanno un loro contenuto empirico sottoposto a metodi di valutazione, dal punto di vista metodologico, molto rigorosi, eccetera, la parte pregiata dei risultati di ricerca finisce in articoli, e il libro è una di quelle operazioni di sintesi che si fanno a volte nella vita accademica, ma non troppe volte, dopo un lungo percorso di ricerca, non prima. La sensazione è che in certe discipline si obblighino ancora i giovani, dopo il dottorato, a produrre comunque una monografia, purché sia, per i successivi passaggi concorsuali, e che lette da un revisore di onesta professionalità queste diano risultati molto scadenti. Lì io credo che dolorosamente alcune comunità abbiano dovuto prendere atto di questo e che vi sia una riflessione da fare che peraltro è già iniziata» (Intervista

Pure in mancanza di dati complessivi alcuni dati parziali sono stati pubblicati, seppure con caratteristiche del tutto peculiari. L'eccezione è rappresentata infatti dai dati relativi al confronto tra le valutazioni peer e la valutazione bibliometrica (Anvur, 2013d[188]). In appendice al rapporto dell'Area delle Scienze Politiche e Sociali vengono riportati alcuni confronti tra le valutazioni fornite dai revisori, ma qui, come per la maggior parte delle Aree, i dati sono limitati al sottoinsieme di prodotti sottoposti anche a valutazione bibliometrica (261 prodotti in totale, *cfr.* § 4.3.2.1) e mirati al confronto delle classi assegnate tramite i giudizi dei revisori e tramite l'analisi bibliometrica (Anvur, 2013d, GEV14, p. 107).

I confronti tra i giudizi dei revisori sono presentati suddivisi in base alle classi di merito ottenute dai prodotti tramite la valutazione bibliometrica: *eccellente* o *buona* (E+B) oppure *accettabile* o *limitata* (A+L, Tabella 42 e Tabella 43).

Tabella 42 - Confronto tra valutazioni bibliometriche nelle classi E+B e A+L e i giudizi dei due revisori (adattamento delle tabelle A.12 e A.14, Anvur, 2013d, GEV14, pp. 107-108)

Classi E+B						
		Revisore 2				Totale
		E	B	A	L	
Revisore 1	E	19	19	3	1	42
	B	17	27	11	3	58
	A	0	10	2	0	12
	L	1	10	1	0	12
Totale		37	66	17	4	124

Classi A+L						
		Revisore 2				Totale
		E	B	A	L	
Revisore 1	E	7	18	3	3	31
	B	11	24	13	5	53
	A	8	4	3	4	19
	L	3	8	8	5	24
Totale		29	54	27	17	127

I confronti complessivi vengono approfonditi presentando anche tavole distinte, una relativa esclusivamente alle coppie di revisori entrambi con affiliazione straniera l'altra relativa esclusivamente alle coppie di revisori entrambi con affiliazione italiana (Anvur, 2013d, GEV14, pp. 107-108; Tabella 43).

Il commento fornito dal GEV a questi dati è estremamente sintetico e vale la pena riportarlo per esteso: «su un totale di 124 prodotti classificati "Eccellente" o "Buono" dalla valutazione bibliometrica, 86 ricevono la stessa valutazione da entrambi i revisori (66%). La percentuale sale al 76% se i revisori sono entrambi stranieri e scende al 63% se entrambi italiani. […] L'accordo tra la valutazione bibliometrica e i due revisori sui prodotti classificati dalla bibliometria nel versante

Bonaccorsi). Una riflessione che trova riscontro anche nell'intervista al professor Colozzi: «un punto di non ritorno è il fatto di aver cominciato a far capire che il veicolo di comunicazione della ricerca scientifica non è solo la monografia, o non è prevalentemente la monografia. Anzi, che sarebbe bene spostarsi ancora di più dalle monografie agli articoli su rivista, ai saggi su rivista. Anche per il modo con cui adesso si fa ricerca, perché vista l'assoluta mancanza di finanziamenti, […] Cambiare la comunicazione della ricerca, il modo di comunicare il lavoro di ricerca, per cui la monografia è sicuramente importante, la monografia ha un suo significato, nel nostro settore non credo sia eliminabile. […] Noi continueremo a scrivere delle monografie a carattere di ricerca, però non sono il veicolo di eccellenza della comunicazione, il veicolo deve diventare un altro» (Intervista Colozzi).

[188] I dati sono generalmente riportati in appendice e spesso commentati brevemente anche nei rapporti di Area: GEV1; GEV2; GEV3; GEV4; GEV5; GEV6; GEV 7; GEV8, Appendice 3; GEV9; GEV13; GEV 14, Appendice C). Fanno eccezione l'Area 11, il cui GEV ha pubblicato, per SSD, numero e percentuali delle valutazioni concordanti e minimamente divergenti (Anvur, 2013d, GEV 11, p. 42, Tab. 2.17-2.18) e l'Area 12, il cui GEV ha reso note le quote di prodotti con valutazioni concordanti, minimamente divergenti o marcatamente divergenti (Anvur, 2013d, GEV 12, p. 67, Tab. 5.1). Nel caso dell'Area 10 nessun dato è stato reso noto sul grado di accordo delle valutazioni dei revisori.

"bottom" è molto inferiore. Infatti, su un totale di 127 prodotti classificati "Accettabile" o "Limitato" dalla valutazione bibliometrica, in 20 casi entrambi i revisori confermano le classi (16%). Questa percentuale rimane pressoché costante sia nel caso di revisori entrambi stranieri e sia entrambi italiani» (Anvur, 2013d, GEV14, p. 107).

Tabella 43 - Confronto tra valutazioni bibliometriche nelle classi E+B e A+L e i giudizi dei due revisori e affiliazione dei revisori (adattamento delle tabelle A.13, A13.1, A15 e A.16, Anvur, 2013d, GEV14, pp. 107-108)

Classi E+B							Classi A+L						
Entrambi i revisori stranieri													
		Revisore 2				Totale			Revisore 2				Totale
		E	B	A	L				E	B	A	L	
Revisore 1	E	2	3	1	0	6	Revisore 1	E	5	6	0	2	13
	B	6	5	2	0	13		B	5	8	6	2	21
	A	0	1	0	0	1		A	5	1	0	1	7
	L	0	1	0	0	1		L	1	4	2	5	12
Totale		8	10	3	0	21	Totale		16	19	8	10	53
Entrambi i revisori italiani													
		Revisore 2				Totale			Revisore 2				Totale
		E	B	A	L				E	B	A	L	
Revisore 1	E	5	6	0	1	12	Revisore 1	E	2	12	3	1	18
	B	5	9	8	2	24		B	6	16	7	3	32
	A	0	2	0	0	2		A	3	3	3	3	12
	L	0	2	0	0	2		L	2	4	6	0	12
Totale		10	19	8	3	40	Totale		13	35	19	7	74

Le brevi annotazioni del GEV circa le tavole mettono in luce soprattutto il maggior grado di accordo tra i revisori stranieri nelle classi E+B, e il minor grado di accordo nelle classi A+L sia con riferimento alle coppie di revisori italiani che con riferimento alle coppie di revisori stranieri, tuttavia le percentuali sono riferite alle classi aggregate per il confronto con le valutazioni bibliometriche. Calcolando le percentuali di giudizi concordanti (C), discordanti di una sola classe (D1), di due classi (D2) e di tre classi (D3), tenendo conto cella per cella delle classi di discordanza e non considerando insieme E+B ed A+L (Tabella 44), è possibile avanzare osservazioni diverse da quelle del GEV.

Tabella 44 – Schema per l'individuazione dei giudizi concordanti, discordanti di una classe, discordanti di due classi e discordanti di tre classi

		Revisore 2			
		E	B	A	L
Revisore 1	E	C	D1	D2	D3
	B	D1	C	D1	D2
	A	D2	D1	C	D1
	L	D3	D2	D1	C

Nel caso delle classi bibliometriche E+B si evidenzia un grado di concordanza più elevato (38,7% contro 30,7%, Tabella 45), ma le coppie di revisori stranieri non presentano una quota di casi di concordanza più elevata rispetto ai revisori italiani, la differenza è piuttosto nell'assenza di casi di discordanza di tre classi. La differenza evidenziata dal GEV tra revisori italiani e stranieri in riferimento alle classi E+B risulta qui più evidente nella classe A+L, dove i revisori stranieri presentano il 34% di giudizi concordanti, i revisori italiani solo il 28,4%.

Da questi dati non è possibile trarre conclusioni significative a proposito dell'uniformità delle scale di giudizio, le valutazioni discordanti di due o tre classi sono frequenti quasi nella stessa misura tra revisori italiani e revisori stranieri. Tuttavia, come sottolineato anche dal GEV, assumendo la valutazione bibliometrica come riferimento (dunque come criterio di validità) sembrerebbe che le scale di giudizio dei revisori stranieri siano maggiormente coerenti sia con riferimento alle classi E+B che per le classi A+L.

Tabella 45 – Percentuale di giudizi concordanti e discordanti di una, due o tre classi per classi E+B e A+L bibliometriche e sul totale

	E+B			A+L			Totale		
	%	% Stra	% Ita	%	% Stra	% Ita	%	% Stra	% Ita
Concordanti	38,7%	33,3%	35,0%	30,7%	34,0%	28,4%	34,7%	33,8%	30,7%
Discordanti di una classe	46,8%	57,1%	52,5%	45,7%	39,6%	50,0%	46,2%	44,6%	50,9%
Discordanti di due classi	12,9%	9,5%	10,0%	18,9%	20,8%	17,6%	15,9%	17,6%	14,9%
Discordanti di tre classi	1,6%	-	2,5%	4,7%	5,7%	4,1%	3,2%	4,1%	3,5%
Totale	100%	100%	100%	100%	100%	100%	100%	100%	100%

Naturalmente le osservazioni avanzate su questi dati non possono essere estese all'intero esercizio, il sottoinsieme dei prodotti sottoposti sia alla peer review che all'analisi bibliometrica non è infatti, almeno per l'Area 14, assimilabile alla totalità dei prodotti sottomessi per la VQR: si tratta soprattutto di prodotti pubblicati in lingua inglese, in forma di articolo su rivista, e l'insieme presenta una sovra rappresentazione di alcuni SSD (in particolare SPS/04 ed SPS/09, *cfr.* Anvur, 2013d, GEV14, pp. 96-100).

5.2 La rilevazione nella procedura di valutazione bibliometrica nella VQR

La valutazione diretta tramite analisi bibliometrica sembrerebbe porre meno problemi in relazione alla rilevazione, in particolare rispetto alla selezione delle fonti. Le problematiche connesse alla selezione dei revisori sono più che evidenti mentre l'uso di database *ad hoc* è in grado di infondere all'intera procedura un'aura di obiettività impensabile nel caso della peer review[189]. Nonostante questa impressione, le citazioni, come tutti i dati, sono soggette a vari generi di distorsioni e la struttura e gli algoritmi dei database possono influire sensibilmente sulla validità e sull'affidabilità delle informazioni che forniscono.

Bornmann e Daniel (2008) in un'ampia rassegna degli studi sul comportamento citazionale hanno classificato i fattori esterni al discorso scientifico in grado di influenzare la probabilità di essere citati: fattori legati al tempo, al campo di studi, alla rivista, al tipo di documento, agli autori o ai lettori, l'accessibilità della pubblicazione e i problemi tecnici legati alla correttezza delle citazioni hanno tutti effetti più o meno noti e prevedibili sul numero di citazioni e possono rendere opaco il legame tra queste e la qualità scientifica di un prodotto.

[189] La traccia di questa sensazione emerge chiaramente dalle interviste, ad esempio: «se ci riferiamo all'analisi bibliometrica, direi che la metodologia scelta dovrebbe essere affetta marginalmente da problemi di errore, nel senso che l'analisi citazionale si è riferita alla produzione mondiale di articoli nello stesso anno e *subject category*, e la teoria e la pratica bibliometrica indicano che la *subject category* è una unità di analisi appropriata» (Intervista Bonaccorsi).

Qui è di particolare interesse l'ultima classe di fattori: la correttezza delle citazioni. La distorsione infatti in questo caso non è legata neppure indirettamente al discorso scientifico, come invece potrebbe esserlo nel caso delle reti citazionali o della lingua di pubblicazione, ma dipende esclusivamente da errori materiali. In uno studio del 1990, analizzando i riferimenti bibliografici degli articoli di tre riviste mediche, Evans e il suo gruppo hanno individuato un 48% di riferimenti incorretti, che si traducono in citazioni non conteggiate (Evans *et al.* 1990).

E' noto inoltre che gli algoritmi di matching tra i documenti citati e le fonti delle citazioni presentino numerosi errori che possono produrre una perdita di citazioni nei conteggi (van Raan, 2005). Moed (2002) stima che in media il numero delle citazioni non conteggiate nei documenti indicizzati sia pari a circa il 7% delle citazioni conteggiate, ma che in specifici casi la perdita delle citazioni può arrivare al 30%. Questi problemi sono spesso relativi a pubblicazioni con numerosi autori e possono essere dovuti a variazioni ed errori nei nomi degli autori (in particolare per gli autori da nazioni di lingua non inglese), ma anche a pubblicazioni in riviste con differenti sistemi di numerazione dei volumi o degli articoli, a errori relativi al numero della rivista o al numero delle pagine (van Raan, 2005). Non stupisce dunque la conclusione di van Raan: «quando gli indici citazionali sono utilizzati per scopi valutativi tutti questi possibili errori devono essere corretti per quanto possibile[190]» (2005, p. 136).

Il numero di citazioni, così come i vari indici che ne derivano, non possono essere considerati come "misure oggettive" o "non distorte" dell'impatto o della qualità scientifica di una pubblicazione: si tratta di indicatori in senso proprio, cioè di informazioni utilizzabili per la stima di una proprietà non altrimenti rilevabile che tuttavia non rappresentano completamente la proprietà da rilevare e contemporaneamente possono essere riferite ad altre proprietà.

La scelta del database non è indifferente in ambito scientometrico né dovrebbe esserlo nella valutazione della ricerca: le basi di dati non includono gli stessi elementi, coprono in maniera diversa diversi campi disciplinari, la loro struttura è differente, riviste e documenti sono classificati sulla base di schemi diversi e persino gli algoritmi per il conteggio delle citazioni e gli indici calcolati sono dissimili[191].

E' vero che le informazioni fornite dai diversi database risultano abbastanza coerenti, non di rado i risultati degli studi *ad hoc* mostrano una elevata correlazione tra i dati estratti da diversi database in riferimento a singoli Paesi o campi disciplinari, ma non sempre i risultati ottenuti da database diversi risultano convergenti (Adam, 2002; Bar-Ilan, 2008; Bakkalbasi *et al.* 2006; Archambault *et al.* 2008; Archambault *et al.* 2009).

[190] Traduzione dall'originale in lingua inglese.

[191] A questo proposito vale la pena sottolineare che nell'Area delle Scienze Chimiche il GEV avrebbe preferito utilizzare un solo database, e che questo non è stato possibile per ragioni essenzialmente tecnico-burocratiche: «nel GEV 3, inizialmente, Scopus non lo volevamo ma per un errore del CINECA, che ha pubblicato un nostro vecchio bando dei criteri, in cui c'era sia… nel primissimo bando c'erano sia Scopus che WoS, poi i chimici non volevano Scopus e così abbiamo deciso di toglierlo, per mesi abbiamo lavorato ragionando solo su ISI fino a che a un certo punto ci hanno fatto notare, un'università, che c'era anche Scopus, quindi abbiamo dovuto… non potevamo più cambiare i criteri pubblicati… abbiamo cambiato la procedura, e abbiamo scelto la valutazione migliore tra le due. Dato che per i chimici la valutazione più alta in genere è ISI non è cambiato moltissimo» (Intervista Carletti); si tratta di una questione di fiducia nei dati, infatti: «il problema è semplicemente che noi crediamo molto più a ISI Web of Science per la chimica specificamente e quindi avremmo preferito usare solo ISI» (Intervista Barone).

Inoltre in relazione all'analisi bibliometrica vanno affrontate le problematiche metodologiche proprie di tutte le analisi secondarie. Scopus e WoS infatti non sono database pensati e costruiti per valutare la ricerca né con riferimento all'Italia né con riferimento ad altri Paesi. I dati dunque potrebbero non rispondere pienamente agli obiettivi cognitivi di un esercizio come la VQR: essere inadeguati dal punto di vista della qualità (intesa sia come validità che come attendibilità), incompleti, incoerenti o anche solo poco tempestivi; il loro utilizzo può richiedere controlli e aggiustamenti di vario genere.

Sulla base della letteratura saranno discussi il livello di copertura dei database (con riferimento all'Area 3: *WoS* e *Scopus*), la loro struttura, gli algoritmi di calcolo e le classificazioni utilizzate, con una particolare attenzione alle possibili influenze di queste caratteristiche sulla qualità dei dati utilizzati nella VQR. Infatti: «dato che Web of Science e Scopus sono database differenti non è possibile assumere che misure riferite alla stessa rivista nei due database conducano automaticamente a un risultato identico (o almeno molto simile)[192]» (Russeau, 2009, p. 11).

5.2.1 La struttura dei database e la rispondenza dei dati agli obiettivi dell'esercizio

La qualità di un database bibliometrico in relazione al suo scopo originale di fonte informativa sulla letteratura scientifica non è determinata, contrariamente a quanto si potrebbe pensare, dall'ampiezza della base di dati. Bradford (1934) ha dimostrato come sia individuabile un nucleo essenziale di riviste e come la maggior parte della letteratura scientifica rilevante in ciascuna disciplina o tematica specifica venga pubblicato in un numero relativamente ristretto di riviste. I *core journals*, quelli cioè che raccolgono la maggior parte della letteratura rilevante su un dato argomento in un dato momento, tuttavia, non sono sempre gli stessi e cambiano con lo svilupparsi di nuove discipline o tematiche, di paradigmi, procedure o prospettive. E' sulla base di questa evidenza che si fonda la selettività dei database bibliometrici, perfettamente in linea con i loro obiettivi originari meno con nuovi utilizzi, tra cui la valutazione della ricerca.

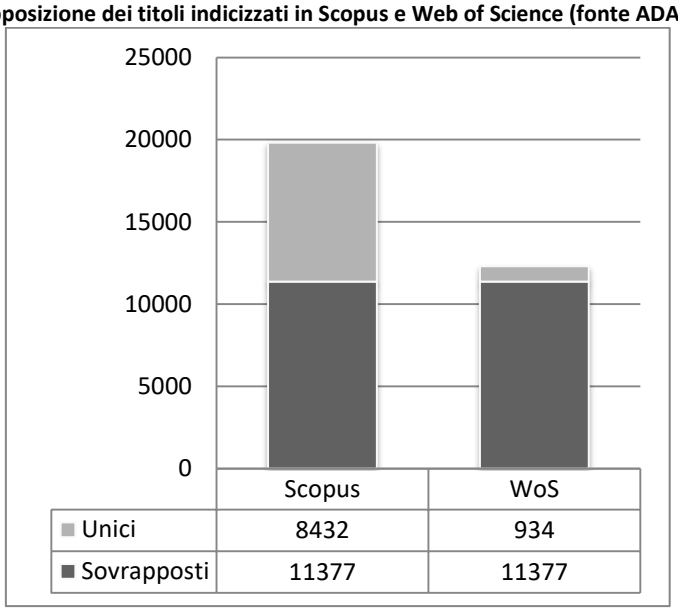
Una prima differenza tra Web of Science di Thomson Reuters e Scopus di Elsevier è individuabile senza dubbio nella loro origine e nelle loro modalità di sviluppo. WoS è stato creato a partire dall'idea di Eugene Garfield (1955) dello *Science Citation Index*, dunque da un nucleo iniziale di *core journals* ampliato via via con l'inclusione nel database di altri titoli in base a una serie di criteri qualitativi e quantitativi (si veda a questo proposito Garfield, 1990). Scopus, invece, è stato creato da Elsevier estraendo i record dai suoi database di abstact (GEOBASE, BIOBASE, ecc..) e aggiungendovi i dati circa le loro citazioni (Jacsó, 2005).

L'approccio è dunque diverso: da un lato WoS è stato creato "dal basso" facendo principalmente riferimento alle citazioni, seguendo l'approccio di Garfield, dall'altro il nucleo essenziale di Scopus è rappresentato dalle riviste inizialmente indicizzate nei database Elsevier, a cui poi si sono aggiunte le altre fonti secondo criteri simili a quelli utilizzati da Thomson Reuters: regolarità, spessore del comitato editoriale, contenuti, rispondenza agli standard editoriali internazionali e citazioni da riviste già indicizzate (Thomson Reuters, 2014; Scopus, 2014).

[192] Traduzione dall'originale in lingua inglese.

E' noto che il numero di fonti (considerando non solo le riviste, ma anche altri generi di documenti come i libri e gli atti di convegno) sia più elevato in Scopus che in WoS (Grafico 3), proprio in ragione della differente filosofia alla base della costruzione dei due database: da un lato l'individuazione dei *core journals* per le singole discipline, dall'altro la costruzione di una base dati il più ampia e interdisciplinare possibile[193]. Entrambe le filosofie hanno punti deboli e punti di forza; qui interessa soprattutto sottolineare che una differente base per il conteggio delle citazioni conduce inevitabilmente a differenze nei conteggi per i singoli documenti (Bakkabasi *et al.* 2008).

Grafico 4 - Livello di sovrapposizione dei titoli indicizzati in Scopus e Web of Science (fonte ADATS 2014[194])



L'introduzione di Scopus nel 2004 ha condotto a una serie di cambiamenti nel panorama della scientometria: «in passato lo *Science Citation Index* di ISI era lo standard nell'analisi citazionale. La copertura dei suoi contenuti era selettiva, e conteneva solo le riviste di maggiore importanza, con un impatto citazionale superiore a una certa soglia minima. Scopus ha una copertura più ampia e indicizza anche riviste con un impatto citazionale più basso e una rilevanza più locale o nazionale. Anche Web of Science di Thomson Reuters sta ampliando la sua copertura di contenuti verso riviste nazionali o regionali. Allo stesso tempo, entrambi gli indici hanno aumentato la loro copertura di atti di convegno e libri. I bibliometrici devono analizzare le conseguenze, sulle loro tecniche e i loro indici,

[193] Un elemento ben noto, che forse ha addirittura favorito la scelta di utilizzare due database: «la decisione di tenere entrambi i database è una decisione fondamentale che il Consiglio Direttivo ha preso per non attribuire vantaggi impropri sia economici ai fornitori commerciali, quindi di natura monopolistica, tenendo aperta la competizione nella fornitura di servizi anche agli atenei, sia per ragioni metodologiche, quindi per non lasciare la valutazione bloccata su scelte fatte a priori su tutte le aree. Di fatto alcune Aree hanno scelto prevalentemente ISI, ma non è un caso che alcune Aree, come l'ingegneria e l'informatica, preferiscano Scopus perché questo indicizza con maggiore intensità *conference proceedings* che non sono riviste ma che in discipline con un forte tasso di accelerazione del tasso di progresso scientifico sono spesso più rilevanti delle stesse riviste, o comunque sono un canale privilegiato perché presentano i risultati in pochi mesi invece che negli anni, tipicamente necessari alla peer review su rivista» (Intervista Bonaccorsi).

[194] *Academic Database Assessment Tool* (ADAT) è un progetto supportato da CRL (*Centre for Reserch Libraries*) e *JISC Collections*, mirato all'informazione sulla comparazione di database bibliografici e bibliometrici; i dati sono aggiornati al 2012.

di questi significativi cambiamenti nella copertura. Gli indicatori dovrebbero essere sviluppati per riflettendo l'effetto delle variazioni nel grado di selettività dei contenuti o dei cambiamenti della copertura nel tempo[195]» (Glänzel e Moed, 2013).

Una problematica troppo spesso sottovalutata o fraintesa riguarda i *bias* linguistici dei database bibliometrici (van Leeuwen *et al.* 2001; Archambault *et al.* 2006). Qui non si fa riferimento solo alla copertura dei database ma anche e soprattutto alla loro capacità di rilevare l'impatto di pubblicazioni in lingue diverse dall'inglese; è infatti noto che le riviste e gli articoli non in inglese hanno indicatori d'impatto più bassi rispetto a riviste e articoli in inglese. La spiegazione è relativamente semplice: dato che la copertura della letteratura in lingua inglese è più ampia le citazioni di riviste in questa lingua hanno una maggiore probabilità di essere rilevate rispetto alle citazioni di riviste in altre lingue.

Questa semplice evidenza conduce alla conclusione che: «gli indicatori devono essere interpretati alla luce dei loro limiti intrinseci, come, in questo caso, gli effetti del linguaggio di pubblicazione[196]» (van Raan, 2005, p. 140). In un esercizio come la VQR è possibile immaginare che le distorsioni legate al linguaggio di pubblicazione nei database finiscano per favorire nell'esito della valutazione gli articoli in lingua inglese rispetto ad articoli dello stesso livello ma pubblicati in altre lingue, tra cui l'italiano stesso.

La più ampia base di dati di Scopus potrebbe comportare un vantaggio da questo punto di vista, anche se ristretto. Leydesdorff e colleghi (2010), facendo riferimento al 2007, riportano una percentuale di riviste pubblicate in inglese pari a circa l'87% in WoS e al 81% in Scopus, cui si aggiungono rispettivamente il 10% e il 17% di riviste multi-linguistiche, la maggior parte delle quali però include l'inglese tra le lingue di pubblicazione.

Nel caso della VQR la diversa composizione linguistica delle riviste indicizzate nei due database costituisce un'ulteriore possibile fonte di distorsione: non solo infatti la lingua di pubblicazione potrebbe influenzare la valutazione finale dei prodotti, ma potrebbe farlo in misura differente a seconda del database, della classe tematica, e così via.

Le differenze tra le due basi di dati sono dunque evidenti, ma non è chiaro quanto le procedure utilizzate per tracciare e conteggiare le citazioni nei due sistemi siano simili. E' noto che in WoS la procedura non è particolarmente accurata (Adam, 2002), le citazioni che presentano anche piccole variazioni nel riportare la rivista possono essere non conteggiate o attribuite erroneamente, e questo può condurre a distorsioni nel conteggio delle citazioni. L'errore inoltre non è distribuito casualmente, ma incide in misura maggiore su specifiche tipologie di riviste, ad esempio quelle che pubblicano articoli con sottotitoli, o di articoli, ad esempio quelli con numerosi autori (van Raan, 2005). Rispetto a quelli di WoS i record di Scopus contengono un'informazione più completa, includendo il titolo per esteso della rivista e di ciascun articolo, dunque in potenza gli algoritmi di conteggio potrebbero essere più precisi (Moed e Visser, 2008). Non è noto però in che misura questa maggiore completezza sia in grado di migliorare l'accuratezza del conteggio delle citazioni, anche perché gli studi condotti sono in genere di piccole dimensioni (ad esempio Meho e Yang, 2007; Adriaanse e Rensleigh, 2013).

Le differenze nella struttura dei database che risultano di maggiore interesse dal punto di vista dell'utilizzo dei dati nella VQR sono essenzialmente riferibili a due delle classificazioni utilizzate

[195] Traduzione dall'originale in lingua inglese.
[196] Traduzione dall'originale in lingua inglese.

in entrambi: la classificazione per tipo di documento, riferita ai singoli documenti contenuti delle riviste indicizzate, e la classificazione tematica delle riviste[197].

La classificazione per tipi di documento è differente nei due database sia per i criteri utilizzati che per l'esito che ne deriva. WoS classifica i documenti pubblicati in base a diversi fattori (la presenza di alcune parole o frasi nel titolo, il numero di citazioni effettuate, le pratiche editoriali della rivista, il formato, ecc.), e le principali categorie includono: articoli (il genere usuale di articolo originale, report o saggio), lettere (contributi dai lettori legati a pubblicazioni precedenti), note (commenti tecnici più brevi di un articolo e con uno scopo più mirato), rassegne (articoli di rassegna o indagine sulla letteratura già pubblicata) (Moed, 1996). Scopus classifica i documenti in: articoli, articoli in corso di pubblicazione, conference paper, editoriale, erratum, lettera, nota, rassegna, short survey (brevi presentazioni di ricerche)[198]. L'accuratezza delle classificazioni, soprattutto nel caso di WoS, è stata ampiamente discussa (Harzing, 2013), anche in relazione al fatto che, a seconda delle discipline, i vari tipi di documenti possono assumere un significato e una rilevanza differente, non necessariamente rispettato nel calcolo degli indicatori di impatto (van Leeuwen *et al.* 2007; van Leeuwen *et al.* 2013). In generale nell'analisi bibliometrica, riguardo alla scelta del tipo di documenti da analizzare: «non è semplice prendere una decisione uniforme, che sia appropriata in tutti i casi. Dipende dai campi di studio e dalle domande di ricerca se le lettere o altri tipi di documento devono essere inclusi oppure omessi[199]» (Moed, 1996, p. 181).

In letteratura non sono presenti studi dedicati espressamente al confronto tra la classificazione dei tipi di documento in WoS e quella utilizzata in Scopus (una analisi molto parziale è presentata in Archimboult *et al.* 2009), ma grazie alla collaborazione di Martijn Visser[200], che sta conducendo presso il CWTS di Leiden una serie di ricerche sull'affidabilità dei database citazionali, è possibile confrontarle su un insieme di documenti[201] inclusi in entrambi i database (Tabella 46).

[197] Elemento ampiamente riconosciuto dall'Anvur: «se il punto è guardare alle citazioni presumibilmente un problema attualmente non c'è, nel senso di capacità di raccoglierle dei due database varia molto meno di un tempo. Rimane la questione della classificazioni, le subject category sono diverse quanto questo possa portare a dei problemi o quanto questi siano riconosciuti dalle varie comunità scientifiche come più idonei o meno idonei, questa è una delle questioni» (Intervista Torrini); .

[198] Si veda: info.sciencedirect.com/scopus/scopus-in-detail/content-coverage-guide/metadata.

[199] Traduzione dall'originale in lingua inglese.

[200] Durante la stesura del lavoro di tesi si è avuta occasione di trascorrere un periodo di studio all'estero, della durata di due mesi, presso il CWTS (*Centre for Science and Technology Studies*) dell'Università di Leiden, nei Paesi Bassi. Nel corso di questo periodo si è avuto modo di approfondire le problematiche metodologiche proprie della bibliometria, avvalendosi del consiglio di alcuni tra i maggiori esperti nel campo. Il lavoro è stato seguito principalmente da Thed van Leeuwen, ma la collaborazione di Martjn Visser è essenziale con riferimento a questa sezione dell'analisi.

[201] L'estrazione è riferita al 2012 per entrambi i database, si noti che il matching dei documenti è incompleto dato che la procedura (per una esposizione breve in una diversa applicazione si veda: Moed e Visser, 2008) è ancora in fase di perfezionamento da parte di Martijn Visser, in assenza di un identificativo per i documenti (mentre per le riviste è possibile utilizzare l'ISSN). Questi dati dunque non possono essere letti in relazione alla copertura o alla sovrapposizione tra i database, il loro unico scopo è quello di permettere il raffronto tra le due classificazioni dei documenti.

Tabella 46 - Confronto tra la classificazione dei documenti per tipo in WoS e Scopus (dati estratti da Martijn Visser, CWTS)[202]

		Scopus											Totale WOS
		Article	Article in Press	Business Article	Conference Paper	Editorial	Erratum	Letter	Note	Review	Short Survey	**Totale**	
WoS	Abstract of Published Item	1										**1**	**1**
	Art Exhibit Review	27			3		1		26	27	56	**140**	**1889**
	Article	1142110	4395	67	33939	617	39	1196	5200	44156	2565	**1234284**	**1306294**
	Bibliography	11				1			2	1	1	**16**	**96**
	Biographical-Item	566	89		21	128	1	13	84	44	34	**980**	**4204**
	Book Review	319	37		32	14	8	13	305	245	177	**1150**	**65789**
	Chronology									2		**2**	**2**
	Correction	192	67		2	24	10352	30	40	14	1	**10722**	**12753**
	Dance Performance Review	12							1	9	20	**42**	**412**
	Database Review										1	**1**	**10**
	Editorial Material	17350	413	78	794	25097	27	642	18881	6886	3639	**73807**	**94014**
	Excerpt	1										**1**	**5**
	Fiction, Creative Prose	63						1		75	23	**162**	**298**
	Film Review	40							13	5	35	**93**	**1301**
	Letter	1455	128		12	153	15	34943	1253	61	16	**38036**	**42645**
	Meeting Abstract	270	1		14	19	20	19	95	11	12	**461**	**242751**
	Music Performance Review	8				1			2	4	5	**20**	**9141**
	Music Score										2	**2**	**24**
	Music Score Review										2	**2**	**65**
	News Item	779	8	8	46	15	5	3	3538	95	792	**5289**	**19460**
	Poetry	47	9						11	12	6	**85**	**4797**
	Record Review	5			1	1			2	3	7	**19**	**1845**
	Reprint	57			11	1			3	18	1	**91**	**201**
	Review	13690	230		850	25	2	4	59	53517	1134	**69511**	**73194**
	Script								1			**1**	**8**
	Software Review	20		1						6		**27**	**44**
	Theater Review	16			2				10	63		**91**	**334**
	TV Review, Radio Review	1				1			3	1		**6**	**381**
	Totale	**1177040**	**5377**	**154**	**35727**	**26097**	**10470**	**36864**	**29529**	**105255**	**8529**	**1435042**	**1881958**
Totale SCOPUS		**1598007**	**88129**	**1357**	**402599**	**47914**	**13681**	**46395**	**60331**	**143302**	**20060**	**2421775**	

[202] In verde sono evidenziate le celle in cui le classificazioni risultano perfettamente concordanti.

La classificazione dei documenti per genere, essendo basata su criteri diversi e costituita da un diverso numero di tipi[203], dà esiti diversi nei due database. Ad esempio, 1.142.110 elementi sono classificati come articoli in entrambi ma la classificazione di Scopus concorda con WoS nel 97% dei casi, la classificazione di WoS concorda con Scopus nel 92,5% dei casi (nell'insieme di sovrapposizione per Scopus gli articoli sono 1.177.040 e per WoS 1.306.294; Tabella 46). Inoltre le discrepanze variano a seconda del tipo di documento, ad esempio per le review la classificazione di Scopus concorda con WoS nel 50,8% dei casi, la classificazione di WoS concorda con quella di Scopus nel 72,9% dei casi. Archimboult e colleghi avevano già sottolineato, con riferimento a un sub-set di documenti pubblicati su *Science*, come la differenza nella classificazione degli articoli sia tutto sommato scarsa, mentre la differenza per le review, le lettere e gli editoriali risulti più rilevante (Archimboult *et al.* 2009).

Questa differenza nella struttura dei database sarebbe di scarso interesse in questa sede, se non fosse che sia l'*impact factor* che lo SJR vengono calcolati considerando solo le citazioni provenienti da alcuni tipi di documenti (al denominatore e/o al denominatore, si veda il § 4.1.3; Moed e van Leeuwen, 1996). A questo si aggiunge l'uso di distribuzioni cumulative empiriche distinte, nel caso delle citazioni, separando «gli articoli "scientifici" da quelli di rassegna» (Anvur, 2013d, GEV3, Appendice, p. 24). Date le differenze nelle classificazioni dei due database non è possibile escludere che le valutazioni dei singoli documenti possano variare a seconda del database di riferimento proprio in ragione del genere a cui erano ricondotti in ciascuno di essi.

Neppure la classificazione tematica delle riviste è identica nei due database: WoS include 170 *subject categories* in *Science*, 50 in *Social Science* e 25 in *Art and Humanities*, mentre l'ASJC di Scopus include 308 classi, raggruppate in 27 macro-categorie[204]. Entrambe le classificazioni dunque si fondano su due livelli e in entrambi i database una rivista può essere assegnata a una o più categorie (Gómez-Núñez *et al.* 2014).

La classificazione per Wos viene effettuata dallo stesso gruppo responsabile della selezione sulla base di criteri "euristici" (Pudvkin e Garfield, 2002), sulla base dei titoli e delle citazioni di ciascuna rivista. In Scopus l'assegnazione delle riviste alle categorie veniva inizialmente effettuata sulla base della descrizione dell'elemento; il gruppo di ricerca di SCImago ha introdotto delle modifiche a questo sistema sulla base dello scopo della rivista e dei feedback dei suoi editori, ma neppure queste modifiche hanno influito sulla solidità della classificazione, considerata ancora poco coerente e fruibile (per tutti Jacsó, 2013; Gómez-Núñez *et al.* 2011 e 2014).

Al di là delle critiche alle singole classificazioni tematiche connesse al numero, all'ampiezza o all'adeguatezza delle categorie, si può immaginare quanto la soluzione del problema sia complessa. La letteratura sulla classificazione e la mappatura della scienza è vasta e varia (ad esempio si vedano

[203] Le distribuzioni singole dei documenti per tipo, anch'esse cortesemente fornite da Martijn Visser sono riportate nell'Appendice B (Tabella 1 e Tabella 2). Si noti a questo proposito che non tutti i tipi sono inclusi nell'insieme di sovrapposizione, restano esclusi gli Abstract report di Scopus e alcuni tipi di review di WoS.

[204] La scelta della categoria tematica non sembra rappresentare un problema dal punto di vista degli EV, che tuttavia a distanza di due anni non riescono a ricostruire esattamente tutti i passaggi, mentre dal punto di vista dei membri dell'Agenzia, nonostante la consapevolezza delle differenze, emerge una forte fiducia nelle procedure: «la teoria e la pratica bibliometrica indicano che la SC è una unità di analisi appropriata. Naturalmente c'è un problema legato alle pubblicazioni multidisciplinari, così dette, per le quali era stato immaginato anche un percorso di peer review complementare all'analisi bibliometrica, però in ogni caso era il soggetto a selezionare la SC in cui essere valutato, per cui da questo punto di vista credo che problemi di qualità del dato non ci fossero» (Intervista Bonaccorsi).

Glänzel e Shubert, 2003; Janssens et al, 2008; Rafols e Leydesdorff, 2009; Zhang *et al.* 2010; Waltman e van Eck, 2012), ma nessuno degli approcci proposti, basati essenzialmente su tecniche di *clustering*, costituirebbe una soluzione definitiva per l'individuazione e soprattutto la definizione delle categorie tematiche, dato che questa questione è strettamente connessa agli obiettivi delle singole indagini (descrittivi, esplorativi, valutativi, più o meno centrati su singole discipline o aree disciplinari, ecc.).

A proposito della VQR in rapporto alla classificazione tematica delle riviste è opportuno considerare una serie di questioni centrali: in che misura le classificazioni di WoS e Scopus sono sovrapponibili? Le riviste risultano classificate in categorie semanticamente affini nei due database? Quando lo sono, la stessa rivista ottiene la stessa classe di merito se considerata in una SC o in una classe dell'ASJC?

Nonostante sia evidente che una stessa rivista possa occupare posizioni diversi nei ranking relativi a diverse categorie tematiche (Amin e Mabe, 2000), la letteratura su questi temi è esigua e i contributi sono centrati su singole tematiche o discipline (si vedano, ad esempio, Lopez-Illescas *et al.* 2008 per l'oncologia; Abrizah *et al.* 2013 per la biblioteconomia). Le analisi disponibili indicano una sovrapponibilità limitata dei due database e, circa le misure di impatto, evidenziano una maggiore generosità di Scopus, riconducibile al maggior numero di titoli coperti in questo database (Lopez-Illescas *et al.* 2008; Abrizah *et al.* 2013).

Un'analisi approfondita delle problematiche connesse alla classificazione tematica delle riviste va oltre gli obiettivi di questo lavoro di tesi, tuttavia, allo scopo di fornire delle evidenze a supporto delle argomentazioni avanzate, verrà presentata un'analisi esplorativa, condotta su un numero limitato di SC e ASJC selezionate come casi-studio.

Tra le 18 *subject categories* ricondotte dall'Anvur all'Area delle Scienze Chimiche (Anvur, 2013c, pp. 84-92) sono state selezionate quattro categorie, in base:

- alla loro ampiezza, cioè al numero di riviste in esse contenute (nel 20% più basso o più alto della distribuzione cumulata del valore in tutte le 18 SC);
- al loro impatto, cioè al valore dell'IF mediano della categoria (nel 20% più basso o più alto della distribuzione cumulata).

I dati sono stati estratti dal JCR e dal SJR in riferimento al 2012, e includono esclusivamente le riviste dotate di ISSN. Le categorie selezionate sono:

- *Chemistry, multidisciplinary*: alto numero di riviste, basso *impact factor*;
- *Chemistry, physical*: alto numero di riviste, alto *impact factor*;
- *Electrochemistry*: basso numero di riviste, alto *impact factor*;
- *Spectroscopy*: basso numero di riviste, basso *impact factor*.

Una volta selezionati i quattro casi-studio in WoS sono state individuate le categorie corrispondenti, in base alle etichette semantiche, nel ASJC di Scopus (Tabella 16).

Tabella 47 - Etichetta delle categorie, indice di impatto e numero delle riviste in WoS e Scopus[205]

Web of Science	*Categoria*	*Mediana IF*	*Riviste*	Scopus	*Categoria*	*Mediana SJR*	*Riviste*
	Chemistry, multidisciplinary	1,385	152		Chemistry, miscellaneous	0,334	229
	Chemistry, physical	2,115	135		Chemistry, physical and theoretical	0,557	124
	Spectroscopy	1,706	43		Spectroscopy	0,618	42
	Electrochemistry	2,144	26		Electrochemistry	0,661	25

Evidentemente le due categorie con un numero elevato di riviste rappresentano campi disciplinari ampi: uno relativamente uniforme (*Chemistry, physical*), l'altro eterogeneo (*Chemistry, multidiciplinary*), mentre le due categorie con un numero modesto di riviste corrispondono a campi disciplinari specifici, legati a settori e argomenti specialistici.

Le dimensioni della categoria in termini di numero di riviste e il valore mediano degli indici di impatto risultano confrontabili tra le due classificazioni, anche se con alcune eccezioni. In particolare la categoria *Chemistry, miscellaneous* in Scopus risulta più ampia del corrispettivo in WoS, inoltre l'impatto mediano delle riviste incluse nella categoria *Spectroscopy* in Scopus è meno basso di quello delle altre categorie e di quello delle riviste in *Chemistry, physical and theoretical* meno elevato, se li si confronta con i corrispettivi IF mediani in WoS.

Thomson Reuters fornisce una breve descrizione per ciascuna *subject category*, mirata a rendere conto del suo contenuto[206], ma non esiste nulla di simile per le categorie dell'ASJC in Scopus. Non è dunque possibile confrontare le definizioni delle categorie disciplinari messe a punto dai due database, è tuttavia di un certo interesse considerare le definizioni delle *subject categories* di WoS.

> **Chemistry, Multidisciplinary:** «include fonti con un approccio generale o interdisciplinare alla scienze chimiche. Anche le fonti su specifici argomenti che hanno una rilevanza per molte aree delle scienze chimiche sono incluse in questa categoria. Le fonti con un focus privilegiato sulla chimica analitica, inorganica e nucleare, organica, fisica o sui polimeri sono classificate nelle loro specifiche categorie[207]».

> **Chemistry, Physical:** «include fonti sulla fotochimica, sulla chimica dello stato solido, cinetica, catalisi, chimica quantistica, chimica di superficie, elettrochimica, termodinamica chimica, termofisica, chimica dei colloidi, fullereni e zeoliti».

> **Electrochemistry**: «copre le fonti che hanno a che fare con i cambiamenti chimici prodotti dall'elettricità e dalla generazione di elettricità da reazioni chimiche. Le applicazioni includono pile asciutte, lamine di piombo, accumulatori, galvanostegia, elettrolisi, purificazione del rame, produzione di alluminio, celle a combustibile e corrosione dei metalli».

[205] Elaborazione propria, su dati estratti dal JCR di WoS e dal SJR di Scopus per l'anno 2012. Così per tutte le analisi presentate in questo paragrafo.

[206] Le definizioni sono consultabili all'interno del JCR, ISI-Web of Knowledge, alla pagina: admin-apps.webofknowlegde.com/JCR/static_html/scope_notes/SCIENCE/2012/SCOPE_SCI.htm.

[207] Traduzione dall'originale in lingua inglese, così per le altre definizioni.

Spectroscopy: «copre le fonti legate alla produzione, misurazione e interpretazione degli spettri elettromagnetici derivanti da emissione o assorbimento di energia radiante da varie fonti. Questa categoria include le fonti che rendicontano circa le diverse tecniche per l'analisi degli spettri delle onde di particelle o per determinare lo spettro di massa».

Le definizioni delle categorie lascerebbero prevedere, almeno all'interno di WoS, una completa separazione di *Chemistry, multidiciplinary* da *Chemistry, physical*, una certa sovrapposizione tra la prima e *Spectroscopy* ed *Electrochemisty*. Inoltre, stando alla sua definizione, la categoria *Chemistry, physical* dovrebbe presentare una sovrapposizione con *Electrochemistry.*

L'obiettivo in questa sede non è il confronto della copertura dei due database, ma l'analisi del livello di sovrapposizione tra le classificazioni delle riviste utilizzate in ciascuna di esse. A questo fine, utilizzando i codici ISSN e i titoli estesi come chiavi, sono stati individuate le riviste presenti in ciascuna coppia di categorie, sia all'interno dello stesso database sia nell'intersezione tra i due database (Tabella 48). Le frequenze più elevate corrispondono all'incrocio delle coppie di categorie semanticamente affini nei due diversi database.

Tabella 48 – Sovrapposizione delle categorie, numero di riviste classificate per coppie di categorie (WoS e Scopus)

		Scopus				*WoS*			
		Misc.	**Phys.**	**Spectro.**	**Electro.**	**Multi.**	**Phys.**	**Spectro.**	**Electro.**
Scopus	**Misc.**	229	10	1	1	114	14	2	1
	Phys.		124	2	6	6	68	7	2
	Spectro.			42	1	-	7	24	-
	Electro.				25	1	3	-	18
WoS	**Multi.**					152	9	1	-
	Phys.						135	7	2
	Spectro.							43	-
	Electro.								26

Impiegando il numero di "riviste uniche" (calcolato come il totale delle riviste meno il numero delle riviste in entrambe, cioè meno i "doppi") incluse nelle due categorie per ciascun incrocio come "massimo della sovrapposizione possibile" è possibile calcolare l'entità, in termini percentuali, della sovrapposizione tra le categorie (Tabella 49). *Electrochemistry* è la categoria con la sovrapposizione più elevata tra i due database, seguita da: *Chemistry, multidisciplinary Spectrocopy* e *Chemistry, physical*.

E' interessante notare che mentre all'interno di Scopus la sovrapposizione più consistente è individuabile tra le aree *Chemistry, physical and theoretical* ed *Electrochemistry*, in WoS la sovrapposizione maggiore è tra *Chemistry, physical* e *Spectrocopy*, sebbene nella definizione della prima vi sia un riferimento esplicito all'elettrochimica. Nonostante quanto riportato nella sua descrizione la categoria *Chemistry, multidiciplinary* in WoS non risulta completamente indipendente da *Chemistry, physical*, né la sua sovrapposizione con le due specialità *Spectroscopy* ed *Electrochemisty* risulta consistente.

E' inoltre possibile notare che la categoria *Chemistry, physical* è l'unica tra quelle di WoS a presentare una sovrapposizione con ciascuna delle categorie di Scopus, mentre sia *Chemistry, miscellaneous* che *Chemistry, physical and theoretical* di Scopus hanno almeno una rivista in comune con ciascuna delle quattro categorie WoS in analisi.

Tabella 49 – Sovrapposizione delle categorie, percentuale del massimo della sovrapposizione possibile (WoS e Scopus)

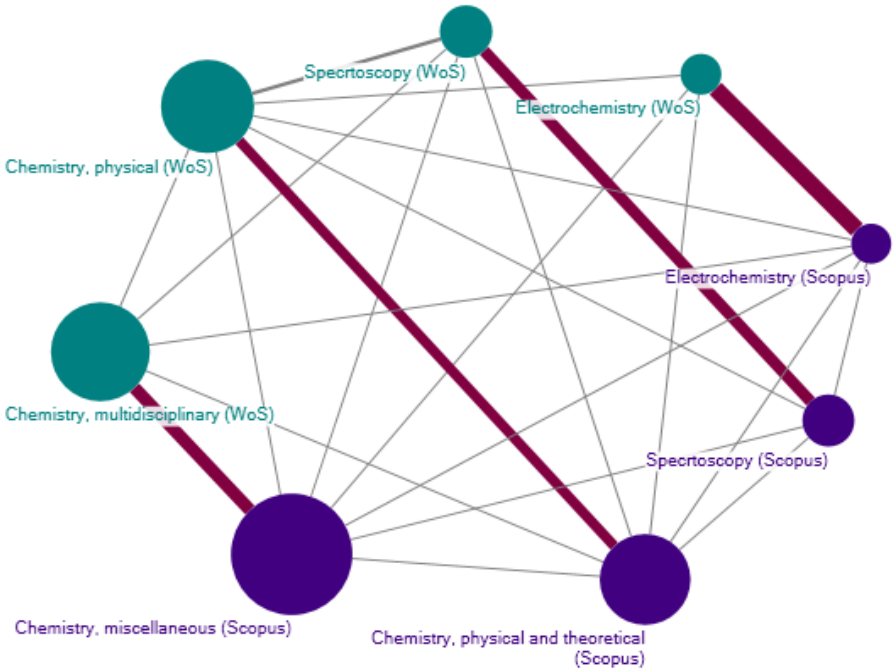
		Scopus				*WoS*			
		Misc.	**Phys.**	**Spectro.**	**Electro.**	**Multi.**	**Phys.**	**Spectro.**	**Electro.**
Scopus	**Misc.**	100,0%	2,9%	0,4%	0,4%	42,7%	4,0%	0,7%	0,4%
	Phys.		100,0%	1,2%	4,2%	2,2%	35,6%	4,4%	1,4%
	Spectro.			100,0%	1,5%		4,1%	39,3%	
	Electro.				100,0%	0,6%	1,9%		54,5%
WoS	**Multi.**					100,0%	3,2%	0,5%	
	Phys.						100,0%	4,1%	1,3%
	Spectro.							100,0%	
	Electro.								100,0%

Concependo le categorie come nodi in una rete e il numero di riviste in comune come legami tra categorie, è possibile ottenere una rappresentazione grafica della sovrapposizione tra le categorie, in cui tanto i nodi quanto i legami siano proporzionali alle loro dimensioni (Grafico 5).

Evidentemente vi è un certo grado di sovrapposizione tra le due classificazioni, ma non si può dire che sia elevatissimo, almeno in riferimento alle categorie selezionate come casi studio. La procedura utilizzata nel corso della VQR non prevedeva regole standardizzate per la selezione della categoria in cui valutare il prodotto. Si è già sottolineato che nell'Area delle Scienze Chimiche, in ogni caso in cui il prodotto risultasse classificato in più di una SC (o ASJC), è stata presa in considerazione l'indicazione fornita dal soggetto valutato, che tuttavia se necessario poteva essere modificata dal GEV (Anvur, 2013d, GEV3, p. 90). Inoltre, la scelta della categoria era completamente indipendente nei due database.

E' dunque necessario considerare le conseguenze che la scelta dell'uno o dell'altro database, e/o dell'una o dell'altra categoria, può avere sul risultato della valutazione.

Grafico 5- Rappresentazione in forma di rete della sovrapposizione tra categorie in WoS e Scopus



Ripercorrendo i passi dell'Anvur, per ciascuna categoria sono state calcolate le distribuzioni cumulate di IF o SJR e le soglie corrispondenti alle classi Anvur (Tabella 50), allo scopo di assegnare a ciascuna rivista un livello di merito sulla base della sua posizione nel ranking.

Tabella 50 – Soglie per l'attribuzione delle classi di merito per categoria (WoS e Scopus)

		SJR				***IF***			
		Misc.	Phys.	Spectro.	Electro.	Multi.	Phys.	Spectro.	Electro.
80-100%	*Soglia tra Buono e Eccellente*	0,853	1,068	1,211	1,301	4,304	4,920	3,947	3,141
60-80%	*Soglia tra Accettabile e Buono*	0,444	0,762	0,876	0,919	1,772	2,434	2,672	2,142
50-60%	*Soglia tra Limitato e Accettabile*	0,374	0,557	0,701	0,681	1,392	2,145	2,279	1,855

Ai fini della classificazione in classi di merito la categoria di riferimento non è indifferente neppure all'interno dello stesso database: ad esempio in Scopus una rivista classificata sia in "*Chemistry, physical*" che in "*Electrochemistry*" con un SJR pari a 0.6 otterrebbe la classe Accettabile se valutata nella prima categoria, la classe Limitato se valutato nella seconda. In WoS una rivista classificata sia in "*Chemistry, multidisciplinary*" che in "*Chemistry, physical*" con un IF di 1.9 otterrebbe la classe Buono se valutato nella prima categoria, Limitato se valutato nella seconda.

Non si tratta esclusivamente del valore dei due indici: nella determinazione delle classi di merito non influisce esclusivamente il valore dell'IF o del SJR, ma anche il sottoinsieme delle riviste ordinate per quel valore. In altri termini la posizione di una rivista nel ranking dipende sia dal suo impatto che da quello delle altre riviste classificate nella stessa *subject category* o ASJC.

Nonostante infatti in relazione alle posizioni nei ranking il coefficiente di cograduazione di Spearman (Tabella 51) risulti elevato, la diversa ampiezza delle categorie, e dunque delle classi al loro interno, potrebbe avere delle conseguenze sulla classe di merito finale. E' per questa ragione che la scelta della categoria andrebbe considerata con la massima attenzione.

Tabella 51 – Rho di Spearman, per le posizioni nei ranking delle riviste per SC e ASJC (*=sig. <0,05; **=sig. <0,01)

			Scopus				*WoS*			
			Misc.	**Phys.**	**Spectro.**	**Electro.**	**Multi.**	**Phys.**	**Spectro.**	**Electro.**
Scopus	**Misc.**	Rho	1	1,000**			,954**	,893**	1,000**	
		N	229	10	1	1	114	7	2	1
	Phys.	Rho		1	1,000**	1,000**	,886*	,929**	,857*	1,000**
		N		123	6	2	6	68	7	2
	Spectro.	Rho			1			,964**	,928**	
		N			42	1	1	7	24	
	Electro.	Rho				1		1,000**		,959**
		N				24		3		18
WoS	**Multi.**	Rho					1	1		
		N					80	2		
	Phys.	Rho						1	1,000**	1
		N						69	4	2
	Spectro.	Rho							1	
		N							24	
	Electro.	Rho								1
		N								17

La prima questione è: una stessa rivista ottiene la stessa classe di merito in categorie semanticamente simili nei due diversi database? L'analisi a coppie delle categorie selezionate come

casi studio permette di individuare alcune risposte e di metterle in relazione con le principali caratteristiche delle categorie stesse: ampiezza e impatto mediano.

I valori dei due indici in riferimento alle riviste incluse sia in *Chemistry, multidisciplinary* che in *Chemistry, miscellaneous* risultano correlati linearmente (Grafico 6), tuttavia le posizioni nei rispettivi ranking e di conseguenza le classi di merito ottenute dalla stessa rivista nei due database presentano delle discrepanze (Grafico 7).

Grafico 6 - IF e SJR per *Chemistry, multidisciplinary* (WoS) e *Chemistry, miscellaneous* (Scopus)

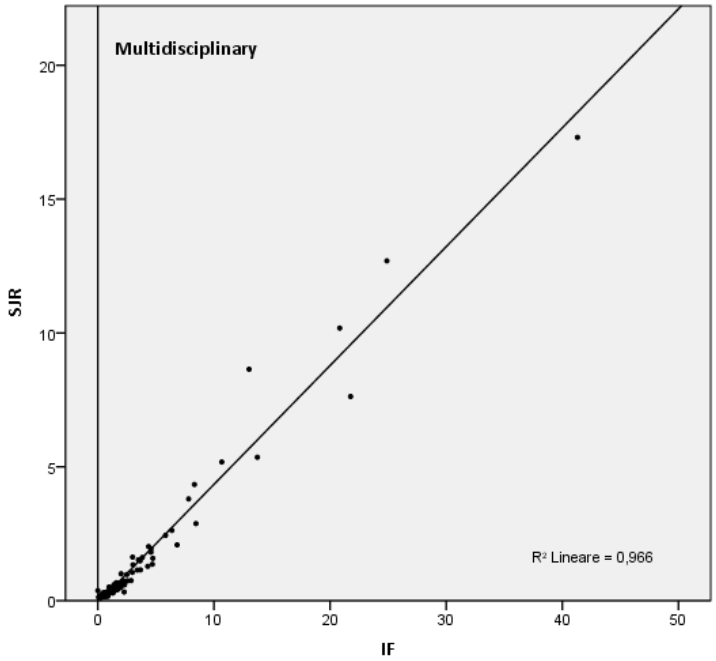
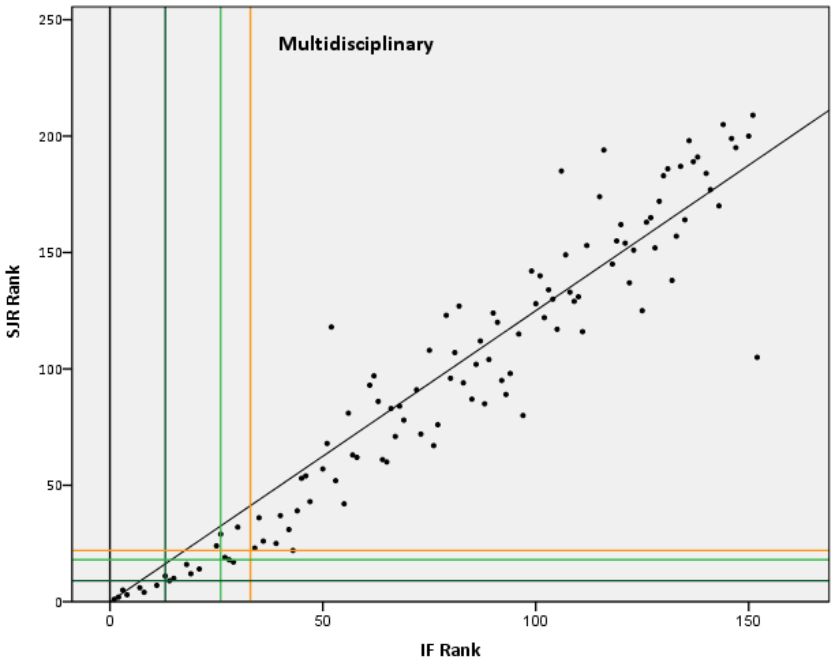


Grafico 7- Posizioni nei ranking per IF e SJR per *Chemistry, multidisciplinary* (WoS) e *Chemistry, miscellaneous* (Scopus), proiezione delle soglie Anvur per le classi di merito



Confrontando le classi di merito ottenute dalle riviste presenti sia in WoS che in Scopus e classificati in *Chemistry, multidisciplinary* e *Chemistry, miscellaneous* è possibile evidenziare che in quasi il 30% dei casi le classi risultano differenti (Tabella 52 e Tabella 53). Inoltre solo per due riviste la classe di merito ottenuta in WoS è più elevata di quella ottenuta in Scopus.

Tabella 52 –Classi di merito nel JCR e nel SJR per le riviste classificate in *Chemistry, multidisciplinary* (WoS) e *Chemistry, miscellaneous* (Scopus)

		Classe di merito IF				Totale
		E	B	A	L	
Classe di merito SJR	E	19	10			29
	B		8	9	6	23
	A			2	7	9
	L		1	1	51	53
Totale		19	19	12	64	114

Nel 6,4% dei casi le valutazioni ottenute dalla stessa rivista risultano discordanti di due classi di merito: sei riviste "Limitate" in WoS risultano "Buone" in Scopus, una "Limitata" in Scopus risulta "Buona" in WoS (Tabella 53)[208].

Tabella 53 – Casi di concordanza e discordanza nell'assegnazione delle classi di merito per le riviste classificate in *Chemistry, multidisciplinary* (WoS) e *Chemistry, miscellaneous* (Scopus) (valori assoluti e percentuali)

	Totale		Più alto Wos		Più alto Scopus	
	n	%	n	%	n	%
Concordanti	80	70,2%				
Discordanti di una classe	27	23,7%	1	0,9%	26	22,8%
Discordanti di due classi	7	6,1%	1	0,9%	6	5,3%
Totale	114	100,0%				

IF e SJR risultano linearmente correlati anche in riferimento alle riviste incluse nelle aree *Chemistry, physical* e *Chemistry, physical and theoretical* (Grafico 8), ma osservando le posizioni nei ranking risulta evidente la differenza tra i due database: le stesse riviste ottengono posizioni più elevate in Scopus che in WoS (Grafico 9).

[208] In Appendice si riporta l'elenco completo delle riviste indicizzate in entrambe le categorie, corredate di indici di impatto, posizione nel ranking e classe di merito (Appendice B, Tabella 3).

Grafico 8 - IF e SJR per *Chemistry, physical* (WoS) e *Chemistry, physical and theoretical* (Scopus)

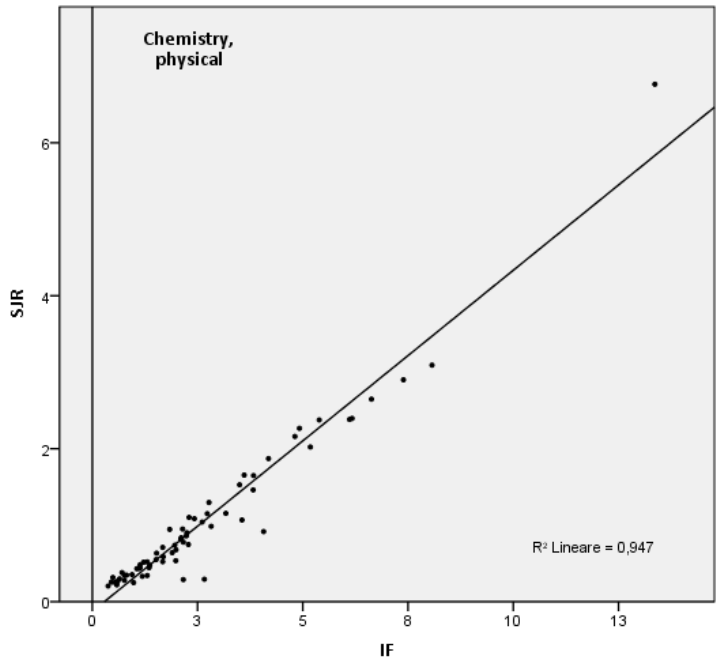
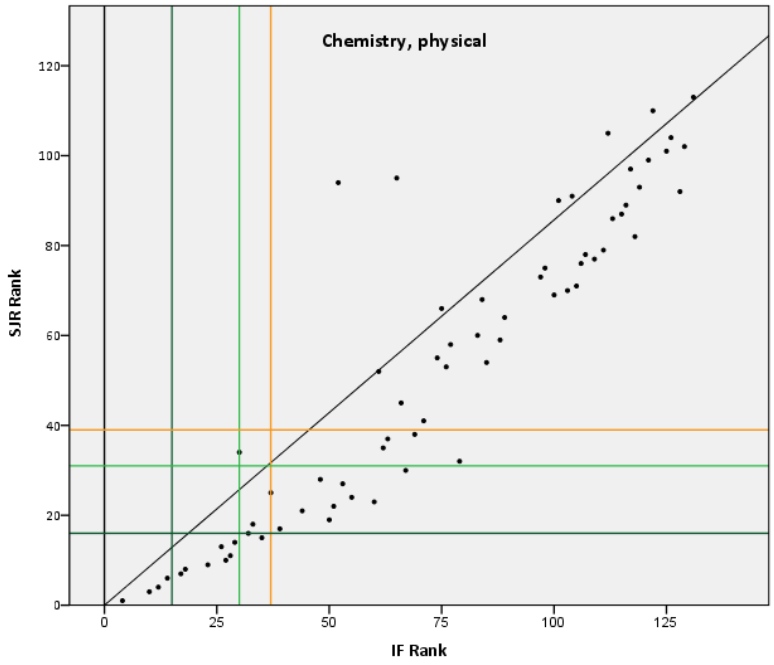


Grafico 9 - Posizioni nei ranking per IF e SJR per *Chemistry, physical* (WoS) e *Chemistry, physical and theoretical* (Scopus), proiezione delle soglie Anvur per le classi di merito



In questa categoria quasi il 40% delle valutazioni risulta discordante, di nuovo in soli due casi la classe di merito ottenuta sulla base dell'IF risulta più elevata di quella ottenuta grazie al SJR, mentre il 29,4% ottiene una valutazione più elevata di una classe e l'8,8% di due classi in Scopus rispetto a WoS (Tabella 54 e Tabella 55)[209].

[209] L'elenco completo delle riviste in *Chemistry, physical* e *Chemistry, physical and theoretical* è riportato in appendice, corredato di indici di impatto, posizione nel ranking e classe di merito (Appendice B, Tabella 4).

Tabella 54 - Classi di merito nel JCR e nel SJR per le riviste classificate in *Chemistry, physical* (WoS) e *Chemistry, physical and theoretical* (Scopus)

		Classe di merito IF				Totale
		E	B	A	L	
Classe di merito SJR	E	9	10	2		21
	B		3	4	3	10
	A			1	6	7
	L		1	1	28	30
Totale		9	14	8	37	68

Tabella 55 - Casi di concordanza e discordanza nell'assegnazione delle classi di merito per le riviste classificate in *Chemistry, physical* (WoS) e *Chemistry, physical and theoretical* (Scopus) (valori assoluti e percentuali)

	Totale		Più alto Wos		Più alto Scopus	
	n	%	n	%	n	%
Concordanti	41	60,3%				
Discordanti di una classe	21	30,9%	1	1,5%	20	29,4%
Discordanti di due classi	6	8,8%	1	1,5%	6	8,8%
Totale	68	100%				

Nell'intersezione delle categorie *Spectroscopy*, la relazione lineare tra IF e SJR è leggermente più debole, verosimilmente anche in relazione alla minore ampiezza del numero di riviste interessate (Grafico 10). La minore ampiezza delle due categorie permette anche di visualizzare meglio le differenze nel ranking (Grafico 11).

Grafico 10 - IF e SJR per *Spectroscopy* (WoS) *e Spectroscopy* (Scopus)

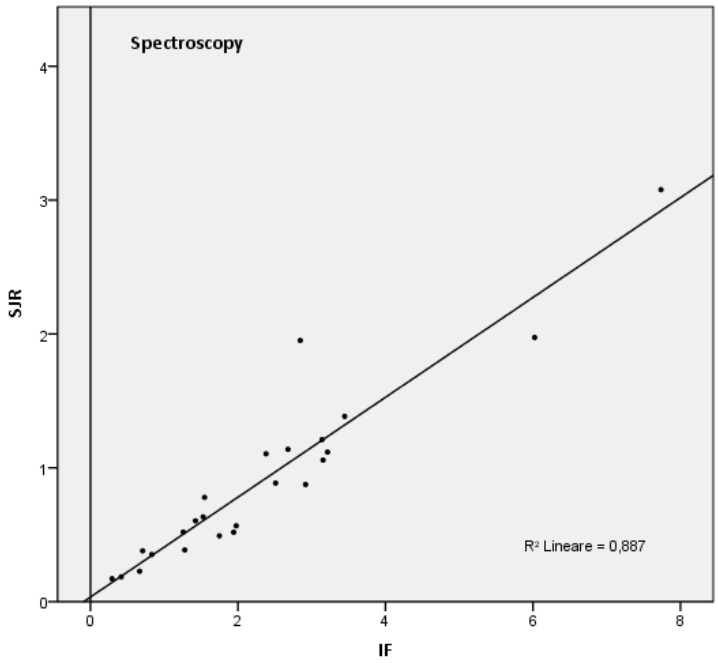
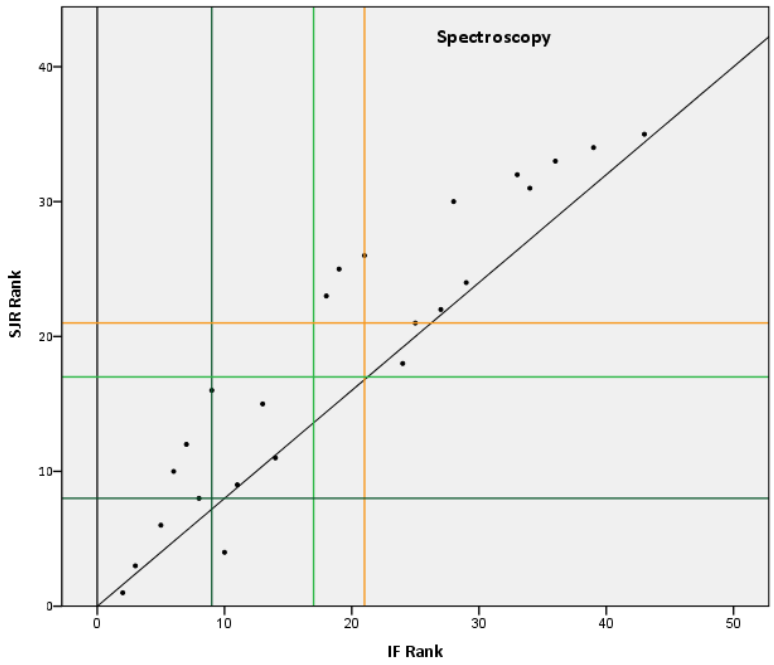


Grafico 11 - Posizioni nei ranking per IF e SJR per *Spectroscopy* (WoS) *e Spectroscopy* (Scopus), proiezione delle soglie Anvur per le classi di merito



Il 25% delle valutazioni risulta discordante, ma tutte discordano di una sola classe (Tabella 56 e Tabella 57), inoltre nel 16,7% dei casi la classe di merito più elevata è quella di Scopus, nell'8,3% quella di WoS[210].

Tabella 56 - Classi di merito nel JCR e nel SJR per le riviste classificate in *Spectroscopy* (WoS) *e Spectroscopy* (Scopus)

		Classe di merito IF				Totale
		E	B	A	L	
Classe di merito SJR	E	4	1			5
	B	2	4			6
	A				1	1
	L			2	10	12
Totale		6	5	2	11	24

Tabella 57 - Casi di concordanza e discordanza nell'assegnazione delle classi di merito per le riviste classificate in *Spectroscopy* (WoS) *e Spectroscopy* (Scopus) (valori assoluti e percentuali)

	Totale		Più alto Wos		Più alto Scopus	
	n	%	n	%	n	%
Concordanti	18	75,0%				
Discordanti di una classe	6	25,0%	2	8,3%	4	16,7%
Discordanti di due classi						
Totale	24	100%				

Le ultime categorie selezionate come casi studio vanno sotto l'etichetta di *Electrochemistry*, e dal punto di vista del numero delle riviste sono le più piccole in analisi. Tra gli IF e i SJR delle riviste presenti in entrambe queste categorie, come negli altri casi, sussiste una relazione lineare (Grafico 12) e anche osservando le posizioni dei ranking cui questi due indici danno luogo è possibile notare

[210] L'elenco completo delle riviste in *Spectroscopy*, con le informazioni circa il valore degli indici di impatto, la posizione nel ranking e la classe di merito Anvur è riportato in appendice (Appendice B, Tabella 5).

che la distanza dei casi dalla bisettrice (cioè dall'eguaglianza delle posizioni) è meno ampia rispetto a tutte le altre categorie (Grafico 13).

Grafico 12- IF e SJR per *Electrochemistry* (WoS) e *Electrochemistry* (Scopus)

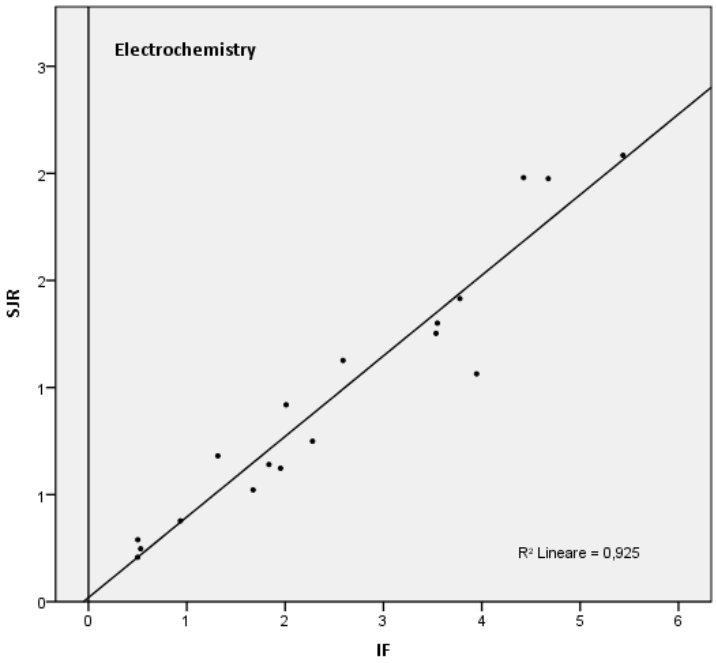
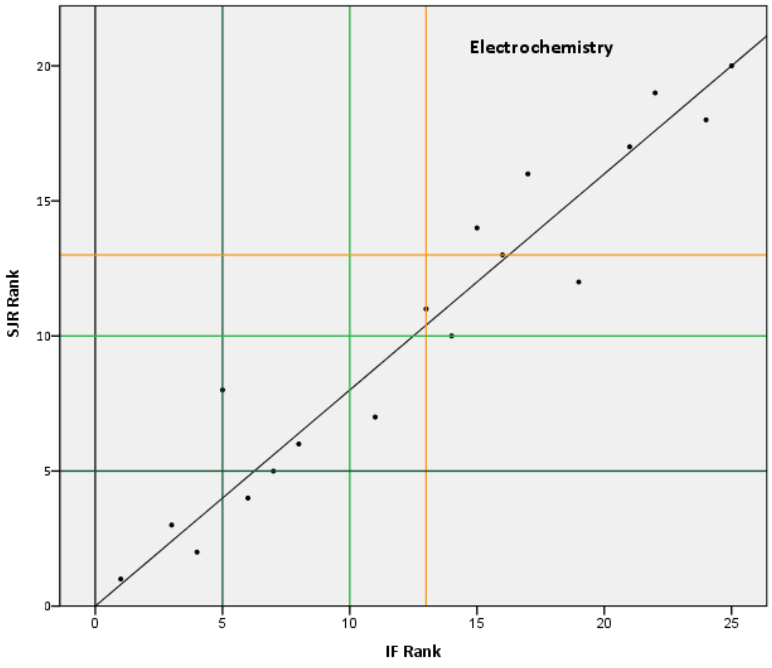


Grafico 13 - Posizioni nei ranking per IF e SJR per *Electrochemistry* (WoS) e *Electrochemistry* (Scopus), proiezione delle soglie Anvur per le classi di merito



Nonostante la minore discrepanza tra i ranking anche nelle categorie di *Electrochemistry* le valutazioni discordanti sono oltre il 30%, di nuovo le valutazioni di Scopus risultano più generose di

quelle di WoS (di una sola classe per il 22,2% dei casi, di due per un solo caso), con una sola eccezione (Tabella 58 e Tabella 59)[211].

Tabella 58 - Classi di merito nel JCR e nel SJR per le riviste classificate in *Electrochemistry* (WoS) e *Electrochemistry* (Scopus)

		Classe di merito IF				Totale
		E	B	A	L	
Classe di merito SJR	E	3	2			5
	B	1	1	1	1	4
	A			1	1	2
	L				7	7
Totale		4	3	2	9	18

Tabella 59 - Casi di concordanza e discordanza nell'assegnazione delle classi di merito per le riviste classificate in *Electrochemistry* (WoS) e *Electrochemistry* (Scopus)(valori assoluti e percentuali)

	Totale		Più alto Wos		Più alto Scopus	
	n	%	n	%	n	%
Concordanti	12	66,7%				
Discordanti di una classe	5	27,8%	1	5,6%	4	22,2%
Discordanti di due classi	1	5,6%			1	5,6%
Totale	18	100%				

In tutte le categorie la valutazione ottenuta in Scopus risulta nella maggior parte dei casi più migliore di quella ottenuta in WoS. Una possibile spiegazione a questa evidenza potrebbe essere legata alle diverse politiche di inclusione dei due database. E' in effetti plausibile che la maggiore generosità delle valutazioni di Scopus sia dovuta all'estensione della sua copertura (in altri termini la maggiore presenza di riviste da aree geografiche e disciplinari periferiche potrebbe spingere in alto nella distribuzione cumulata le riviste internazionali indicizzate anche in WoS).

Una seconda questione da affrontare in relazione alle categorie disciplinari è: una stessa rivista ottiene la stessa classe di merito in categorie semanticamente diverse, nello stesso database o in due diversi database?

La quota di valutazioni concordanti per coppia di categorie (Tabella 60) è, quasi sempre, elevata tra le coppie di categorie nello stesso database.

Tabella 60 – Casi con classi di merito concordanti negli insiemi di sovrapposizione tra categorie

		Scopus				*WoS*			
		Misc.	**Phys.**	**Spectro.**	**Electro.**	**Multi.**	**Phys.**	**Spectro.**	**Electro.**
Scopus	**Misc.**	229 (100%)	6 (60%)	0 (0%)	0 (0%)	80 (70,2%)	8 (57,1%)	1 (50%)	0 (0%)
	Phys.		124 (100%)	2 (100%)	5 (83,3%)	3 (50%)	41 (60,3%)	1 (14,3%)	0 (0%)
	Spectro.			42 (100%)	1 (100%)		4 (57,1%)	18 (75%)	
	Electro.				25 (100%)	1 (100%)	1 (33,3%)		11 (61,1%)
WoS	**Multi.**					152 (100%)	7 (77,8%)	0 (0%)	
	Phys.						135 (100%)	5 (71,4%)	2 (100%)
	Spectro.							43 (100%)	
	Electro.								26 (100%)

[211] Naturalmente anche l'elenco delle riviste incluse nelle categorie *Electrochemistry* è riportato in appendice (Appendice B, Tabella 6), insieme alle informazioni circa il valore degli indici di impatto, la posizione nel ranking e la classe di merito Anvur.

In WoS (Tabella 60) la concordanza supera il 70% in tutte le coppie, con l'eccezione di quella tra *Chemistry, multidicisciplinary* e *Spectroscopy,* che include una sola rivista. In Scopus (Tabella 60) *Chemistry, miscellaneus* è la categoria che presenta la più bassa quota di casi concordanti: il 60% con *Chemistry, physical and theoretical* e nessuno con *Spectroscopy* ed *Electrochemistry* (con cui tuttavia ha in comune una sola rivista).

Le quote di valutazioni concordanti tra categorie semanticamente diverse di diversi database sono in genere meno elevate: superano il 50% in tre casi, di cui uno con una sola rivista inclusa in entrambe le categorie coinvolte.

La scelta della *subject category* da utilizzare è dunque in grado di introdurre delle distorsioni, modificando l'esito della procedura di valutazione in direzioni non prevedibili ex-ante[212]. Sarebbe opportuno uno studio approfondito delle due classificazioni utilizzate, del loro livello di sovrapposizione e delle conseguenze sugli esiti della valutazione, con l'obiettivo di fissare delle regole standard, almeno a livello di Area, per l'assegnazione dei prodotti alle *subject categories*[213].

L'analisi appena condotta fa riferimento esclusivamente alle riviste, ma è il caso di sottolineare che esiti simili, se non amplificati, possono essere immaginati in riferimento ai ranking degli articoli per numero di citazioni. Il numero di articoli incluso in ciascuna categoria è, logicamente, molto più elevato del numero di riviste e dipende, almeno in parte, dalle caratteristiche delle riviste incluse nella categoria; inoltre, il numero di citazioni può variare fortemente tra un articolo e l'altro, e anche la sua variabilità dipende fortemente dalla rivista e dal campo disciplinare.

La struttura dei database non offre dunque sicurezze circa la validità e l'attendibilità dei dati né al livello dell'individuazione delle citazioni né al livello delle classificazioni sulla cui base vengono calcolati (nel caso del *document type*) o confrontati (nel caso delle *subject categories*) gli indici di impatto, né le distorsioni che possono dipendere da questi elementi risultano controllabili o prevedibili. E' chiaro che la responsabilità di queste distorsioni è attribuibile ai gestori dei database e che ben poco può essere fatto dall'esterno per migliorare le classificazioni e gli algoritmi in uso, nondimeno le agenzie di valutazione possono (e dovrebbero) proporsi con forza come stakeholders, al fianco della comunità scientifica e degli esperti nel campo, nel richiedere ai gestori lo sviluppo di database e strumenti sempre più affidabili: «nessuno dovrebbe accettare una macchina di valutazione a scatola nera[214]» (Hicks *et al.* 2015).

---

[212] Il fatto che sia il soggetto valutato a scegliere la categoria nelle intenzioni dell'Agenzia costituisce una garanzia: «i criteri dei GEV sono pubblicati prima. Quindi i candidati che devono sottomettere i prodotti ai GEV sanno qual è la metrica e ormai gran parte degli Atenei sono in grado di fornire un servizio di previsione citazionale. Quindi credo che nel prossimo esercizio, ora che tutti i ricercatori sono avvezzi a queste procedure, sapranno come calcolarsi le citazioni su entrambi i sistemi e quindi anche su *subject categories* diverse, e lo potranno fare alla luce dei criteri del GEV. Il punto è che dando il GEV i criteri a priori non dà dei vantaggi a nessuno ed essendo calcolati gli indicatori sulla media della propria area, questo sembra equo, nel senso che rende maggiormente non comparabili i giudizi su aree diverse, non solo bibliometriche e non bibliometriche, ma anche tra diverse bibliometriche. Quindi dal punto di vista metodologico insistiamo sulla non comparabilità, però invece verticalmente sull'area sono corrette. Una volta che è pubblicato il criterio non si vedono obiezioni» (Intervista Bonaccorsi). D'altro canto, la capacità di sfruttare questa garanzia dipende almeno in parte dalla preparazione e dalle risorse degli autori e delle strutture.

[213] Ad esempio con una procedura di riconduzione simile a quella utilizzata per i prodotti pubblicati in riviste *multidisciplinary science* e per i prodotti i cui autori non avevano segnalato alcuna categoria, cioè sulla base della categoria prevalente tra le riviste citate nel prodotto.

[214] Traduzione dall'originale in lingua inglese.

Riguardo alle fonti di dati secondarie, una questione sempre centrale e spesso problematica è la tempestività della disponibilità dei dati stessi. I dati bibliometrici, tuttavia, rendono la questione ancora più complessa: un dato articolo o una data rivista possono infatti essere indicizzate a poca distanza dalla pubblicazione, ma i dati relativi alle loro citazioni potrebbero non essere "maturi" per un lasso di tempo fortemente variabile a seconda dei campi di studio o anche del genere di documento (van Raan, 1996). Apparentemente, un confronto all'interno dei singoli campi di studio (pure con le imperfezioni delle classificazioni tematiche in uso) è in grado di limitare il rischio di distorsioni connesse al tempo trascorso dalla pubblicazione, ma vanno considerati altri elementi.

In un lavoro recentissimo Campanario (2015) ha evidenziato che il contributo maggiore all'*impact factor* per il 41% delle riviste proviene dagli articoli pubblicati 2 o 3 anni prima dell'anno di riferimento. Una buona scelta è rappresentata da una finestra di almeno tre anni, una sorta di compromesso per cogliere sia la ricezione veloce di discipline come la medicina e la biologia sia la recezione meno immediata che caratterizza altri campi di studio (si veda ad esempio: Moed *et al.* 1995; van Raan, 1996; Glänzel, 2008)[215].

Nuovamente, i problemi relativi agli indici di impatto utilizzati per la VQR 2004-2010 si estendono, e in questo caso si amplificano, con riferimento al conteggio delle citazioni. Nell'Area delle Scienze Chimiche i dati estratti da WoS e Scopus erano infatti aggiornati al 31 dicembre 2011, e i conteggi delle citazioni includevano le citazioni ricevute dalla pubblicazione a quella data. La differenza tra le finestre temporali può costituire una fonte di distorsione (*cfr.* § 4.1.3), ma qui ci interessa mettere in risalto il fatto che al momento della realizzazione dell'esercizio, tra l'aprile 2012 e il luglio 2013, i dati più aggiornati possibile erano proprio i dati aggiornati al 31 dicembre 2011, utilizzati nella maggior parte delle Aree (con la sola eccezione dell'Area 13).

L'impossibilità di valutare ex ante la maturità dei dati citazionali e la necessità di utilizzare dati validati pongono un evidente dilemma: tanto più una valutazione bibliometrica è tempestiva (cioè il più vicina possibile nel tempo al periodo di riferimento) tanto meno sarà attendibile (cioè tanto meno saranno i campi di studio e le riviste con un livello di maturità citazionale tale da permetterne effettivamente una valutazione per via bibliometrica)[216].

[215] Data l'impossibilità di fissare una finestra utile per tutti i campi di studio vale la pena di citare brevemente la proposta di Dorta-González and Dorta-González (2013), riferita all'impact factor. L'indice in questione prevede di considerare una finestra di due anni che però non sia cronologicamente fissa come per l'attuale impact factor, ma che sia mobile in modo tale che possa catturare il massimo impatto per quella rivista (l'indice viene denominato *2-year maximum journal impact factor*). Il discorso può essere senza dubbio esteso anche allo SJR.

[216] Di nuovo vale la pena riportare le parole del professor Bonaccorsi: «nelle discipline dove il tempo di citazione può essere maggiore di due o tre anni, quando l'esercizio è ogni quattro, potrebbe esserci un problema di rappresentatività delle citazioni, quindi credo che da questo punto di vista i GEV dovrebbero fare una attenta riflessione sulla esclusività dell'analisi bibliometrica soprattutto per i prodotti degli ultimi due anni. La letteratura mostra che negli ultimi due anni è meglio l'impact factor, nel senso che predice meglio la storia citazionale futura, rispetto alle citazioni primissime, perché i lavori che arrivano ad essere molto citati in un anno o due sono pochi e sono di qualità altissima, ma non è detto che non essere citati nei primi due anni sia di per sé un indicatore di bassa qualità. Spesso il picco di citazioni avviene negli anni successivi» (Intervista Bonaccorsi).

Conclusioni

Il nodo centrale in riferimento alla rilevazione dei dati è lo stesso già evidenziato in relazione alla definizione operativa strettamente intesa: l'assenza di trasparenza e argomentazione di alcune scelte essenziali ai fini dell'affidabilità dei dati prodotti dall'esercizio.

La selezione dei revisori e l'assegnazione dei prodotti nella procedura di valutazione tramite peer review sono passaggi cruciali, eppure nei rapporti dell'Agenzia lo spazio dedicato a questi aspetti è estremamente ridotto. In particolare sarebbe stata opportuna la pubblicazione delle informazioni aggregate circa l'aderenza dei profili dei revisori ai criteri stabiliti dai GEV, le loro caratteristiche rilevanti, i carichi di lavoro (il numero e la varietà dei prodotti revisionati), senza contare le informazioni relative all'assegnazione dei prodotti e dunque alla distribuzione congiunta delle caratteristiche rilevanti di revisori, prodotti e autori. Una analisi del grado di accordo tra i revisori inoltre, avrebbe potuto guidare la riflessione sulla costruzione degli standard di qualità all'interno delle comunità scientifiche, rappresentando un investimento per il futuro[217], oltre che risultare di grande interesse dal punto di vista della sociologia della scienza.

Nella procedura di valutazione diretta tramite analisi bibliometrica la questione è la scelta dei database: in diverse Aree di più database. Nonostante l'Anvur argomenti questa scelta e i singoli GEV riportino argomentazioni identiche a quelle proposte dall'Agenzia è evidente una sottovalutazione del problema. L'uso dei dati bibliometrici, così come è in tutte le analisi di dati secondari, avrebbe potuto essere accompagnato dalla riflessione sulle caratteristiche dei database, sulle classificazioni in uso, sulla loro accuratezza e correttezza. E' evidente che i tempi imposti dall'esercizio avrebbero escluso in ogni caso un intervento di controllo sulle basi di dati, nondimeno sarebbe stato possibile presentare le principali criticità dei dati, anche solo a latere dei rapporti. Da un lato è vero che si tratta di problematiche che interessano soprattutto gli addetti ai lavori, dall'altro in nome del principio di accountability sarebbe stata opportuna la loro presentazione proprio a vantaggio dei soggetti meno addentro alla materia. SI tratta cioè anche qui di una questione di trasparenza: i soggetti coinvolti nell'esercizio dovrebbero essere informati circa la natura e il grado di affidabilità dei dati utilizzati, anche e soprattutto perché, nel caso della VQR, potrebbero tenerne conto nella selezione dei prodotti da inviare a valutazione.

Non è da sottovalutare il fatto che la messa in evidenza da parte dell'Agenzia dei limiti dei database disponibili potrebbe costituire un mezzo di pressione, nei confronti dei fornitori, nella direzione del miglioramento dell'accuratezza e della affidabilità dei dati.

[217] Nella direzione indicata da Bonaccorsi: «bisogna fare un enorme lavoro sulla peer review, per persuadere le comunità conflittuali che la peer review è in grado di valutare onestamente anche i lavori di comunità ostili dal punto di vista metodologico, culturale, religioso, politico, tutte quelle variabili che possiamo mettere in gioco. Bisogna credo insistere su questa linea, appunto, che potremmo chiamare habermassiana, di un terreno, che mi viene da chiamare democratico, di costruzione di consensi democratici ragionati dentro le comunità scientifiche» (Intervista Bonaccorsi).

Capitolo 6

Alcune proposte per la valutazione dei prodotti della ricerca

Introduzione

L'obiettivo dell'analisi metodologica presentata non è una critica sterile della valutazione della ricerca, ma lo sviluppo ed il miglioramento delle procedure in uso. Lo scopo è infatti quello di proporre e di permettere in un prossimo futuro l'implementazione di protocolli valutativi meno esposti ad alcuni tra i rischi individuati e dunque in grado di rilevare dati sempre più validi e attendibili. Sono avanzate in questa sezione del lavoro alcune proposte mirate a ridurre questi rischi.

Il concetto di qualità della ricerca non è messo in discussione nella elaborazione delle proposte, così come ha rappresentato il punto di riferimento essenziale per l'analisi delle procedure. Un prima ragione è che modificando il concetto da rilevare la corrispondente definizione operativa non risulterebbe confrontabile con quella utilizzata nel corso della VQR 2004-2010. Inoltre la normativa prevede che la definizione dei criteri non sia operata dall'Agenzia, ma dal Ministero dell'Istruzione, dell'Università e della Ricerca, contestualmente all'indizione dell'esercizio di valutazione. Nonostante criteri e classi di merito risultino eccessivamente vaghi e ambigui rispetto ai loro scopi (*cfr.* Capitolo 3) è necessario tenere conto delle definizioni in uso, pur segnalando i passaggi in cui una minore vaghezza e ambiguità favorirebbero la messa a punto di definizioni operative più affidabili.

Le proposte relative a definizioni operative più attente alle questioni della validità e dell'attendibilità e a procedure di rilevazione e selezione delle fonti mirate alla riduzione delle distorsioni saranno dunque calibrate sulla base del concetto di qualità della ricerca così come definito dal Ministero.

6.1 La definizione operativa della qualità

Alla luce dell'analisi presentata nel Capitolo 4 la definizione operativa della qualità della ricerca rappresenta il nodo cruciale per la rilevazione di valutazioni affidabili tanto con riferimento alla procedura di valutazione tramite peer review quanto con riferimento alla procedura di valutazione diretta tramite analisi bibliometrica.

Sono infatti la validità e l'attendibilità della definizione operativa a determinare la qualità dei dati rilevati in termini di rispondenza alle condizioni logiche e metodologiche definibili a monte a partire dagli obiettivi cognitivi dell'indagine. In questo caso sono le definizioni operative della qualità della ricerca utilizzate a determinare la rispondenza delle singole valutazioni agli obiettivi cognitivi dell'esercizio di valutazione e alla definizione di qualità della ricerca delineata a monte dal Ministero.

Le proposte avanzate con riferimento alla valutazione in peer review sono riferite essenzialmente alla scheda di valutazione e mirano a ridurre i rischi legati alla formulazione dei criteri, alla scala di valutazione, e alla sintesi dei giudizi.

Nel caso della procedura di valutazione diretta tramite analisi bibliometrica, un primo focus circa le proposte è la riconduzione degli indicatori alle classi di merito, considerando anche le problematiche connesse alla calibrazione della procedura. Una serie di altri aspetti sono legati alla possibilità di migliorare ulteriormente la comparabilità dei dati provenienti dai due diversi database, uniformando le definizioni operative degli indicatori relativi all'impatto delle riviste e la classificazione degli argomenti. La questione essenziale, nondimeno, è l'implementazione di una procedura che assegni ai GEV e ai gruppi di consenso un reale ruolo di controllo, in modo tale che l'esito dell'analisi bibliometrica non venga utilizzato senza prima essere stato validato da esperti del campo.

6.1.1 Proposte per la valutazione in peer review

L'analisi metodologica della definizione operativa del concetto di qualità della ricerca utilizzata nella valutazione tramite peer review nella VQR (*cfr*. Capitolo 4) ha evidenziato soprattutto la carenza di trasparenza nei rapporti dell'Anvur. La presentazione delle proposte sarà dunque accompagnata da una serie di argomentazioni relative alla rilevazione dei giudizi dei pari, in particolare alla struttura e ai contenuti semantici della scheda di valutazione, alla scelta delle scale di valutazione, alla riconduzione dei punteggi alle classi di merito per i singoli revisori e alla classe di merito finale.

L'istituzione di un nesso tra il concetto di qualità della ricerca e la formulazione delle domande e delle relative modalità di risposta nella scheda di rilevazione delle valutazioni corrisponde a un rapporto di indicazione. Essendo straordinariamente complicato, ammettendo che sia possibile, selezionare le caratteristiche specifiche in grado di *indicare* la qualità di un prodotto della ricerca (data la varietà delle forme che i prodotti stessi e le loro qualità possono assumere nell'ambito della comunicazione scientifica, nonché la difficoltà nello stabilire gli standard per la valutazione di ciascuna di esse), il parere di un esperto rappresenta il referente empirico più affidabile della qualità; nondimeno gli aspetti da valutare vanno operativizzati in modo tale da permettere una espressione del giudizio che rispetti i criteri di riferimento. Si tratta di una questione centrale dal punto di vista semantico, e determina una serie di scelte connesse allo stesso tempo con la definizione operativa vera e propria del concetto e con la scelta del tipo di variabile in cui l'informazione si trasforma una volta rilevato; una questione forse sottovalutata nel corso della messa a punto degli strumenti per la VQR 2004-2010.

La necessità di ridurre i rischi di distorsione connessi alla formulazione delle domande e delle modalità di risposta conduce a una prima possibile soluzione: l'adozione di uno strumento che risulti il più semplice possibile. E' stato infatti evidenziato, nel corso dell'analisi, come nella scheda di rilevazione utilizzata nell'Area 14 la formulazione delle modalità di risposta presentasse diverse problematiche, in particolare la presenza di più oggetti e la mancanza di mutua esclusività ed esaustività. Queste caratteristiche mettono a rischio la comparabilità dei giudizi espressi e, insieme

alla scelta di proporre solo quattro alternative, potrebbero spingere i revisori verso le modalità di risposta incluse tra gli estremi[218].

Una possibilità è utilizzare i criteri stessi come stimoli, corredati da definizioni che ne chiariscano il significato, richiedendo ai revisori di segnalare un punteggio per ciascun criterio, senza associare ai punteggi etichette semantiche il cui significato possa risultare più o meno condiviso e uniforme tra i revisori (si veda il Riquadro 3).

La scala di valutazione da utilizzare dovrebbe essere sufficientemente ampia da includere una certa varietà di posizioni, senza tuttavia risultare eccessivamente dispersiva; la scelta di una scala a dieci gradienti sembra la più ragionevole. Una scala a nove gradienti infatti, pur essendo sufficientemente ampia (Preston e Colman, 2000) da risultare rassicurante per attendibilità e validità, presenta un gradiente centrale che in qualche modo permette al revisore di trovare un punto di equilibrio per il giudizio sul criterio. La scelta di presentare le scale senza etichettare i gradienti con punteggi numerici è sostenuta dalla letteratura, produce infatti un aumento dell'attendibilità della scala (Cook *et al.* 2001). Tuttavia fornire etichette semantiche ai soli estremi (per nulla/del tutto) su una scala con gradiente centrale, data l'attrattività del gradiente centale, potrebbe dar luogo a più risposte incerte (Cannavò e Basevi, 2003) di quante un esercizio di valutazione con gli obiettivi della VQR potrebbe tollerare. La scala a 10 gradienti è in genere ritenuta più semplice da utilizzare e più adatta all'espressione delle opinioni; in più il suo utilizzo sembra richiedere meno tempo al rispondente rispetto alla scala a nove gradienti (Preston e Colman, 2000, pp. 8-9).

Il fatto che le schede non riportino i punteggi non significa che i revisori debbano esserne all'oscuro, si ritiene infatti che la procedura di rilevazione e sintesi dei giudizi debba essere massimamente trasparente non solo per la comunità, ma anche per i revisori[219]. La formazione e la responsabilizzazione dei revisori peer è infatti il nodo centrale per assicurare validità e attendibilità ai loro giudizi.

Questa scheda (Riquadro 3) è evidentemente analoga alla scheda utilizzata nell'Area 7 (Anvur, 2013d, GEV 7, Appendice p. 28), salvo l'ampiezza della scala di valutazione e il fatto che nella proposta qui avanzata sono previsti diversi campi aperti obbligatori. Nella proposta tre campi aperti sarebbero da destinare ai commenti sui singoli criteri, destinati a una breve motivazione di ciascuno dei giudizi espressi, e uno alle note sul prodotto, da riservare alla registrazione di commenti più generali e appunti destinati agli EV. Le motivazioni dei giudizi e le eventuali note dei revisori renderebbero più semplice il compito degli EV non solo nell'assegnazione della classe di merito finale in casi vicini alle soglie, ma anche nell'individuazione di eventuali fraintendimenti dei criteri di valutazione o, in caso sia opportuno, nella scelta dei terzi revisori.

[218] A questo proposito anche il Presidente del GEV 14 suggerisce alcune modifiche: «siccome l'eccellenza era 9, noi di 9 non ne abbiamo, se non i pochissimi casi di lavori pubblicati su riviste internazionali e valutati di classe A, cioè eccellenti. Quindi abbiamo una sottostima dei lavori eccellenti dovuta a questo *misunderstanding* del concetto di internazionalizzazione. Forse se, oltre a questo, avessimo usato un punteggio un po' più ampio probabilmente avremmo alzato almeno la media, non i prodotti eccellenti ma almeno la media dei prodotti si sarebbe un po' alzata» (Intervista Colozzi).

[219] Circa le diverse scale utilizzate nelle schede il professor Bonaccorsi ha sottolineato che «nella prossima VQR, ed è un mio parere, dico una cosa che non è ancora stata discussa, sarebbe opportuno avere una scala unica perché rende maggiormente confidenti i ricercatori sull'omogeneità del giudizio» (Intervista Bonaccorsi).

Riquadro 3 – Scheda di rilevazione per le valutazioni peer – versione semplificata

D1. ***Rilevanza***

da intendersi come 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 nulla rilevante □ □ □ □ □ □ □ □ □ □ del tutto rilevante

D2. ***Originalità/innovazione***

da intendersi come contributo all'avanzamento di conoscenze o a nuove acquisizioni nel settore di riferimento

per nulla originale □ □ □ □ □ □ □ □ □ □ del tutto originale

D3. ***Internazionalizzazione***

da intendersi come posizionamento nello scenario internazionale, in termini di rilevanza, competitività, diffusione editoriale e apprezzamento della comunità scientifica, inclusa la collaborazione esplicita con ricercatori e gruppi di ricerca di altre nazioni

per nulla internazionale □ □ □ □ □ □ □ □ □ □ del tutto internazionale

Motivazione dei giudizi espressi:

Rilevanza: _____
Originalità:_____
Internazionalizzazione:_____

Note: _____

Indubbiamente questa prima soluzione non è in grado di risolvere le problematiche già evidenziate (*cfr.* Capitolo 4) circa la sotto-determinazione degli stimoli derivante dal riferimento a più oggetti e più attributi. Pur considerando, infatti, come specificazioni del significato i riferimenti alla scienza in generale e alle ricadute nel caso della rilevanza e il riferimento alle nuove acquisizioni per l'originalità, il problema continuerebbe ad avere un certo rilievo nel caso dell'internazionalizzazione, la cui definizione fa riferimento a rilevanza, competitività, diffusione editoriale ed apprezzamento della comunità scientifica, più la collaborazione con ricercatori ed enti stranieri.

Di qui l'individuazione di una seconda soluzione (Riquadro 4), che va oltre la modifica della scheda alla luce dei punti deboli e delle possibili distorsioni individuate nel corso dell'analisi, e parte proprio dalla specificazione del significato dei criteri[220]. La specificazione del significato può essere

[220] Il professor Colozzi ha espresso chiaramente la propria opinione, che vale la pena di riportare, seppure leggermente sintetizzata: «la scheda va assolutamente perfezionata […] Originalità, rilevanza, internazionalizzazione sono i tre criteri. Originalità è… noi sappiamo benissimo che i prodotti veramente originali, dal punto di vista della teoria, sono uno su mille, sono pochissimi […] ovviamente se per originalità si intende che i dati sono originali questo è ovvio, io non posso riciclare come prodotto di ricerca una rimasticatura… è tautologico […] E' ricerca se presento dati nuovi, nel caso di ricerca empirica. Allora cos'è che mi dice la qualità? […] La cosa che importa è il modo in cui i dati sono stati trattati, il modo con cui sono stati presentati, cioè la correttezza metodologica, l'usare le tecniche il più aggiornate possibile, o sofisticate, o

operata individuando dei referenti per i criteri, con lo scopo di chiarire il legame semantico di ciascuna domanda presente nella scheda con il relativo criterio (a titolo di esempio si veda il Riquadro 4). Questa proposta è legata al fatto che «quanto più è frammentato il giudizio, tanto più gli indicatori divengono concreti e tanto meno è chiamata nel giudizio la soggettività del ricercatore; diventa più facile l'accordo inter-soggettivo, aumentano la comunicabilità del procedimento di classificazione e la sua replicabilità» (Agodi, 1999, pp. 129-130).

E' stato più volte sottolineato come la produzione scientifica possa assumere una tale varietà di forme da impedire l'individuazione di caratteristiche fisse cui riferire la qualità, tuttavia i referenti individuati sono di una tale generalità da includere una grande varietà di forme e da permettere la valutazione di qualsiasi prodotto della ricerca. Ciascuno di questi elementi è quasi sempre presente in un prodotto scientifico, seppure con una maggiore o minore centralità rispetto ai suoi obiettivi, e che in ogni caso l'obiettivo fondamentale di un prodotto non può che ricadere in uno di questi ambiti (di seguito saranno discusse anche le conseguenze sulle procedure di sintesi da adottare).

La scheda proposta riporta tre referenti per rilevanza e originalità: il problema di indagine, l'impianto teorico-concettuale e l'impianto metodologico (Riquadro 4). In letteratura con il termine rilevanza si fa spesso riferimento alla rilevanza sociale (ad esempio Scott, 2007), ma qui si intende lasciare al criterio il suo significato più ampio di rilevanza scientifica[221]. Tenendo dunque presente sia la formulazione delle domande nella scheda di rilevazione 2004-2010 sia la definizione utilizzata per il

adeguate, se c'è o non c'è un'interpretazione sociologica, nel caso di un lavoro di tipo sociologico, oppure ci sia un semplice dato statistico. […] Non so come chiamare, diciamo qualità scientifica, un primo criterio potrebbe essere quello della qualità scientifica, piuttosto che l'originalità […] La rilevanza invece è importante, ma la rilevanza a mio parere andrebbe intesa, come noi avevamo suggerito ma non sempre è stato considerato, quanto il lavoro è stato preso o potrebbe essere preso in considerazione dal punto di vista di chi si occupa di quel tipo di problema. […] E' un criterio proponibile quello della rilevanza, quando si lavora con prodotti recenti, freschi, diciamo freschi? Secondo me no, a meno di non dare della rilevanza un'interpretazione che è equivalente a quella di originalità, cioè è una cosa che non si era mai vista prima, ma vale uno su mille. Io penso che anche questo secondo criterio sia difficilmente utile a capire la qualità del lavoro scientifico. Per quanto riguarda l'internazionalizzazione, ho già detto, lì c'è stato proprio… è servita a creare più confusione che altro […] uno non può valutare in termini di referaggio un'ipotesi dell'impossibile, se è scritta in italiano è scritta in italiano, allora cosa vuol dire internazionalizzazione? Non lo so, diventa di fatto un equivalente di rilevanza che diventa un equivalente di originalità. Mi sembra che i criteri utilizzati siano stati un po' confusivi, non abbiano aiutato moltissimo la valutazione e quelli che dovevano fare la valutazione, per cui a mio parere sarebbero interamente da ripensare. Sempre che si vada ancora nella direzione della peer review per la prossima VQR» (Intervista Colozzi). Vanno in questa direzione alcuni propositi del professor Benedetto: «pur essendo inevitabile che ci sia un certo grado di sovrapposizione bisognerà cercare di evitarlo a tutti i costi, per cui la definizione, se ci lasceranno aiutarli a scriverla sarà decisamente più accurata. Oltre alla definizione, ripeto, mentre nel caso degli indicatori bibliometrici problemi non ce ne sono, perché la definizione quantitativa implica la calibrazione e se la calibrazione è fatta bene problemi non ce ne sono, per la peer review oltre ad una definizione più chiara dei criteri bisognerà dare delle informazioni di dettaglio ai revisori» (Intervista Benedetto).

[221] Il criterio relativo alla rilevanza è definito come 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» (Anvur, 2011, p. 7). Tuttavia congruità, efficacia, tempestività e durata sono criteri applicabili solo a una piccola parte dei risultati della ricerca delle Aree in cui la valutazione dei prodotti viene effettuata tramite peer review. Perciò tutti e tre i referenti indivisuati fanno riferimento alla parte fondamentale della definizione, cioè al valore aggiunto per l'avanzamento della conoscenza.

l'esercizio inglese (REF)[222] è evidente quanto la rilevanza scientifica sia essenzialmente riconducibile a una rilevanza sostantiva, una teorica e una metodologica. La formulazione della domanda mira a chiarire ulteriormente il significato del criterio facendo esplicitamente riferimento all'importanza del prodotto per l'avanzamento di conoscenza nel campo.

Circa l'originalità/innovatività, considerando nuovamente la definizione del criterio e la definizione utilizzata per il REF inglese[223], insieme alla formulazione della domanda relativa a questo criterio nella scheda di valutazione di Area 14 è sembrato naturale e opportuno mantenere una simmetria con i referenti individuati per la rilevanza scientifica: il problema d'indagine, l'impianto teorico-concettuale e l'impianto metodologico. La formulazione fa direttamente riferimento al livello di originalità/innovatività del prodotto.

L'internazionalizzazione, invece, nella scheda proposta è rilevata nuovamente senza riferimenti a referenti specifici, ma solo esplicitando il riferimento al posizionamento internazionale del contenuto del prodotto. La scheda di rilevazione di Area 14 faceva riferimento essenzialmente alla visibilità e all'interesse internazionale, tuttavia nella definizione ministeriale si fa riferimento anche alla rilevanza e alla competitività, connotando il posizionamento internazionale non solo in termini editoriali ma anche come contributo al dibattito scientifico. Qui è la stessa formulazione della domanda a porre l'accento sulla connotazione del criterio come posizionamento nello scenario internazionale e la scala di valutazione fa riferimento al grado di centralità del contributo.

La scheda presentata non può che rappresentare un semplice esercizio, dal momento che le definizioni dei criteri e delle classi non permettono di individuare chiaramente i referenti della qualità della ricerca. La proposta avanzata si basa sì sulle definizioni dei criteri e delle classi di merito messe a punto da Ministero e Anvur, ma anche sugli elementi emersi dall'analisi della scheda utilizzata in Area 14 e su considerazioni connesse alla letteratura e all'esperienza internazionale. La proposta dunque, pur essendo direttamente riferita alle Scienze Politiche e Sociali, andrebbe necessariamente discussa entro la comunità scientifica di riferimento.

La scelta dei referenti di ciascun criterio sarebbe da controllare ed eventualmente da calibrare con attenzione. L'adozione di una soluzione di questo tipo dovrebbe prevedere una fase di confronto con la comunità, che sarebbe non solo utile, ma necessaria perché l'intera procedura possa considerarsi affidabile.

Infine, seppure una simile proposta dovesse risultare adeguata allo scopo dal punto di vista della comunità scientifica delle Scienze Politiche e Sociali, dovrebbe essere adeguata e ri-calibrata sulla base delle caratteristiche delle comunità e delle pratiche scientifiche per poter essere utilizzata in altre Aree disciplinari.

[222] «La *rilevanza* sarà intesa in termini di sviluppo dell'agenda intellettuale del campo e può essere **teorica**, **metodologica** e/o **sostantiva**» (REF 01.2012, p. 66-67, Traduzione dall'originale in lingua inglese).

[223] «L'*originalità* sarà intesa nei termini del carattere innovativo del prodotto della ricerca. Un prodotto della ricerca che dimostri originalità può: confrontarsi con **problemi** nuovi e/o complessi, sviluppare **metodi** di ricerca, metodologie e tecniche di analisi innovativi, e/o far avanzare la **teoria** o l'analisi della dottrina, dei criteri o delle pratiche» (REF 01.2012, p. 66-67, Traduzione dall'originale in lingua inglese).

Riquadro 4 – Scheda di rilevazione per le valutazioni peer – versione ampliata

D1. ***Rilevanza***: qual è il livello di importanza del prodotto per l'avanzamento della conoscenza nel campo?

a – Con riferimento al problema d'indagine:
per nulla rilevante □ □ □ □ □ □ □ □ □ □ del tutto rilevante

b – Con riferimento all'impianto teorico-concettuale:
per nulla rilevante □ □ □ □ □ □ □ □ □ □ del tutto rilevante

c – Con riferimento all'impianto metodologico:
per nulla rilevante □ □ □ □ □ □ □ □ □ □ del tutto rilevante

D2. ***Originalità/innovazione***: qual è il livello di originalità/innovatività del prodotto?

a – Con riferimento al problema d'indagine:
per nulla originale □ □ □ □ □ □ □ □ □ □ del tutto originale

b – Con riferimento all'impianto teorico-concettuale:
per nulla originale □ □ □ □ □ □ □ □ □ □ del tutto originale

c – Con riferimento all'impianto metodologico:
per nulla originale □ □ □ □ □ □ □ □ □ □ del tutto originale

D3. ***Internazionalizzazione***: qual è il posizionamento del prodotto nel dibattito internazionale?

per nulla centrale □ □ □ □ □ □ □ □ □ □ del tutto centrale

Motivazione dei giudizi espressi:

Rilevanza: **_____**
Originalità:**_____**
Internazionalizzazione:**_____**

Note: _____

Così come per la scheda semplificata (Riquadro 3), si è scelto di utilizzare una scala di valutazione a dieci gradienti, per le stesse ragioni: offrire una gamma sufficiente di posizioni ai revisori senza rischiare di accentuare distorsioni connesse l'ampiezza della scala. Inoltre le scale auto ancorate forniscono un grado di supporto maggiore alla scelta di utilizzare i punteggi come numeri naturali, pur in assenza della sicurezza che gli intervalli tra i punteggi siano effettivamente percepiti come uguali dai revisori.

La scheda di rilevazione, prima di essere utilizzata nel corso dell'esercizio, dovrebbe essere opportunamente sottoposta a una procedura di pre test, mirata ad identificare eventuali difetti di progettazione e le varie problematiche connesse da un lato all'interpretazione delle domande, dall'altro all'uso delle scale di valutazione da parte dei revisori. A questo fine sarebbe sufficiente una procedura di pre test standard, che preveda, ad esempio, la valutazione di *n* prodotti da parte di *n*

revisori (magari selezionati per quote con un disegno fattoriale che consideri diverse caratteristiche) e un'analisi approfondita dei risultati con particolare attenzione al grado di concordanza e alle eventuali anomalie. Sarebbe tuttavia di gran lunga più interessante un pre test che coinvolga i revisori non solo come rispondenti, ma anche come esperti del campo. Si pensi ad esempio a una procedura di pre test tramite *in depth probes procedure* (per tutti si vedano Lutinski, 1988; Mauceri, 2003), che prevede di far seguire alla compilazione della scheda una serie di domande di approfondimento mirate a controllare tutti quei processi comunicativi e cognitivi che vanno dall'interpretazione dello stimolo alla scelta della modalità di risposta. Questa procedura fornirebbe non solo informazioni sull'adeguatezza dello strumento e sugli eventuali correttivi da apportare, ma anche una serie di opinioni sui criteri, di indicazioni circa le eventuali differenze nella loro interpretazione, evidenze su come gli esperti elaborano un giudizio di qualità in relazione a un prodotto della ricerca. Un pre test di questo genere sarebbe non solo funzionale, ma anche di grande interesse dal punto di vista della sociologia della scienza, la funzionalità comunque è il nodo cruciale. Infatti una tempestiva identificazione e correzione degli eventuali difetti dello strumento si traduce in maggiori garanzie circa l'affidabilità dei dati rilevati e in una riduzione delle problematiche legate alla loro validità. Tanto maggiori sono queste garanzie tanto più fondata sarà l'assegnazione della classe di merito finale al prodotto[224].

Nella procedura utilizzata per la VQR 2004-2010 la sintesi dei punteggi ottenuti su ciascun criterio corrispondeva all'attribuzione di una classe di merito "provvisoria", riferita ai giudizi espressi da un solo *referee*. Questa sintesi, pur avvenendo tramite una operazione matematica, era sostanzialmente basata su considerazioni di ordine semantico e sul rispetto delle quote previste dal bando (Anvur, 2011), tuttavia né la procedura né le considerazioni su cui questa si basava erano esposte nel rapporto finale.

Neppure questa fase è esente da criticità legate, da un lato, al rapporto degli indicatori con il concetto e alle loro relazioni reciproche, dall'altro, al tipo di variabili sintetizzate. L'uso di scale a dieci gradienti rende meno problematica la sintesi dei giudizi per via matematica, e nel rispetto delle definizioni delle classi di merito sarebbe nuovamente pensabile una determinazione delle soglie in linea con le quote 20%-20%-10%-50% previste dal bando.

La procedura 2004-2010 prevedeva la somma dei punteggi e la proposta delle soglie andava chiaramente nella direzione del rispetto delle quote previste dal bando (ad esempio si vadano le soglie per l'assegnazione delle classi in base ai giudizi di due revisori, espressi su schede con punteggi in scala 1-9, Tabella 61)[225]. L'avvicinamento alle quote previste evidentemente era per eccesso piuttosto che per difetto, con la sola eccezione della classe *limitato*.

[224] Una proposta che trova riscontro nelle parole del professor Bonaccorsi: «dovremo fare un esteso lavoro di sperimentazione delle schede per la peer review, chiedendo ai GEV di impegnare del tempo a fare dei pre test, sostanzialmente, per verificarne l'attendibilità. In questa occasione i GEV lo hanno fatto internamente, devo dire con un lavoro anche molto accurato, ma fondamentalmente interno. Questo è un miglioramento che si può introdurre, una fase di pre-test della scheda, basata appunto su test statistici per verificare eventuali distorsioni e per avere anche una base anche di calibrazione che non potrà mai essere perfetta, perché parliamo di numeri anche molto grandi, però che possa migliorare la calibrazione dei giudizi» (Intervista Bonaccorsi).

[225] Infatti circa la determinazione delle soglie è stato possibile rilevare che: «fondamentalmente dovevamo avere la mappatura tra il punteggio e la scala imposta dal bando» (Intervista Bonaccorsi).

La scelta di strutturare le schede con punteggi 1-10 non evita il problema dell'arrotondamento delle quote, anzi la scelta delle soglie implica l'allontanamento delle quote di punteggio corrispondenti alle quote delle classi di merito previste dal bando sia nel caso si scelga di arrotondare per difetto (-) che nel caso si scelga di arrotondare per eccesso (+).

Tabella 61 - Classificazione per due referaggi per le Aree con punteggi in scala 1-9, con le corrispondenti quote di punteggio per classe di merito (adattamento da Anvur, 2014, p. 2) e proposta di classificazione per le scale 1-10

Classe	VQR 2004-2010, scala 1-9		Proposte, scala 1-10			
			-		+	
	Punteggi	Quota corrispondente	Punteggi	Quota corrispondente	Punteggi	Quota corrispondente
Eccellente	23-27	20,8%	26-30	18,5%	25-30	22,2%
Buono	18-22	20,8%	21-25	18,5%	19-24	22,2%
Accettabile	15-17	12,5%	18-20	11,1%	16-18	11,1%
Limitato	3-14	45,9%	3-17	51,9%	3-15	44,5%

Utilizzando la media dei tre punteggi sarebbe possibile presentare soglie teoriche corrispondenti alle quote previste dal bando (Tabella 62), ma la soluzione sarebbe solo apparente. In effetti il numero di possibili punteggi medi riconducibili a ciascuna classe sarebbe pari al numero di somme di punteggi riconducibili a ciascuna classe, con il risultato di ottenere quote teoriche differenti, ma identiche quote di punteggi possibili. Nondimeno le soglie teoriche per la scala 1-10 risulterebbero più intuitive di quelle di qualsiasi altra scala, permettendo ai revisori di tenere a mente le soglie nel corso della valutazione e di avere un'idea immediata e chiara del risultato dei giudizi espressi. Si consideri a questo proposito che la scala 1-10 presenta un parallelo con i voti scolastici che potrebbe contribuire a chiarire il significato delle classi di merito, facendo corrispondere la classe *limitato* ai voti "insufficienti", la classe *accettabile* ai voti che pur essendo al limite della sufficienza non la superano, la classe *buono* ai voti tra il 6 e l'8 e la classe *eccellente* ai voti al di sopra dell'8.

Tabella 62 -Classificazione per due referaggi con punteggi in scala 1-9 e 1-10 sulla base delle medie dei punteggi

Classe	VQR 2004-2010, scala 1-9		Proposta, scala 1-10	
	Punteggio	Quota corrispondente	Punteggio	Quota corrispondente
Eccellente	7,2-9	20%	8,1-10	20%
Buono	5,4-7,1	20%	6,1-8	20%
Accettabile	4,5-5,3	10%	5,1-6	10%
Limitato	1-4,4	50%	1-5	50%

Le sintesi per somma semplice o media dei punteggi presentano però anche altre problematiche. Questa procedura infatti considera ugualmente validi e rilevanti tutti e tre i criteri di qualità della ricerca, e sembra rispondere alla definizione quantitativa delle classi di merito, tuttavia confrontando le definizioni semantiche delle classi di merito riportate dall'Anvur con quelle deducibili dalla scelta delle soglie sono state messe in luce diverse differenze, alcune più rilevanti di altre (*cfr.* § 4.3.2). In sintesi: nelle definizioni *originalità* e *rilevanza* sembrano legati da un legame di co-implicazione e l'*internazionalizzazione* non risulta decisiva per l'assegnazione della classe di merito, mentre la scelta delle soglie sottintende non solo che ciascun criterio assuma lo stesso peso, ma anche che i tre criteri siano ortogonali tra loro.

Queste criticità derivano dall'ambiguità e dalla vaghezza evidenziate in relazione alla definizione dei criteri (*cfr*. § 1.2)[226]. E' stato sottolineato come nelle definizioni manchi il riferimento a qualsiasi regola di classificazione per i prodotti della ricerca con giudizi discordanti sui tre criteri. Nella maggior parte dei casi, dunque, la somma dei punteggi conduce alle classi di merito senza legame con la definizione del concetto. E' il presupposto (tacito) relativo all'ortogonalità dei criteri a permettere di determinare la classe di merito del prodotto tramite soglie: se le tre proprietà da valutare sono indipendenti una dall'altra e connesse in uguale misura al concetto generale di qualità della ricerca è possibile sommare i punteggi rilevati su ciascuna proprietà e assumere che il totale corrisponda alla qualità complessiva del prodotto.

Nella definizione delle classi di merito, però, sono stati individuati diversi elementi che lasciano propendere per una non ortogonalità dei criteri di valutazione (*cfr*. § 3.2). Innanzitutto rilevanza e originalità sembrano legate indissolubilmente, vi è una sovrapposizione parziale tra le loro definizioni e nella definizione delle classi di merito è spesso impossibile individuare chiaramente a quale dei due criteri si faccia riferimento. Inoltre nella definizione di tutte le classi di merito sia presente il riferimento sia al livello nazionale sia al livello internazionale, e l'internazionalità non sembra rappresentare una condizione definitoria necessaria della qualità della ricerca (risultano infatti essenziali originalità e rilevanza)[227].

Al fine di rendere conto della struttura del concetto di qualità della ricerca si propone una procedura di sintesi alternativa. Così come per la seconda proposta di scheda (Riquadro 4) è stato necessario individuare dei referenti per i criteri, sostanzialmente operando una specificazione a partire dalle definizioni ministeriali, qui sarà necessario chiarire e specificare le relazioni tra i criteri, cioè la struttura del concetto di qualità della ricerca.

E' opportuno prima di tutto operare una disambiguazione in relazione ai significati dei criteri *rilevanza* e *originalità*. Nonostante la sovrapposizione parziale delle definizioni non è infatti pensabile che i due criteri siano stati intesi come sovrapponibili dal decisore. Anche nella definizione operativa proposta (Riquadro 4) rilevanza e originalità hanno gli stessi referenti, non per questo però è pensabile che assumano lo stesso significato. L'accento dunque andrebbe posto sulla differenza tra il "valore aggiunto" e il "contributo" all'avanzamento della conoscenza cui si fa riferimento nelle definizioni (Anvur, 2011, p. 7). La definizione delle classi di merito aggiungono alcuni elementi facendo riferimento al «rigore metodologico» e alla «rilevanza interpretativa» oltre che all'originalità

[226] In proposito si veda anche la nota 220, p. 171.

[227] Il Coordinatore dell'esercizio ha in effetti avanzato l'ipotesi che nella prossima VQR i criteri non saranno equivalenti, e che in particolare che: «per la procedura di peer review probabilmente sfumerà un pochino l'importanza del criterio dell'internazionalizzazione in certe Aree, perché è quello che ha creato più problemi, criteri veri e criteri meno veri… per cui tra i tre criteri l'internazionalizzazione non apparirà come uno dei criteri fondanti, ma apparirà in vece in maniera più sfumata […] però è una nostra intenzione, in realtà è una cosa che dovrà scrivere il Ministero nel decreto di apertura della nuova VQR. E' una delle cose che abbiamo indicato al Ministero come punti critici, ma non sappiamo come andrà a finire» (Intervista Benedetto). Se l'internazionalizzazione rappresenta una caratteristica accessoria, e non costitutiva, della qualità scientifica sarebbe anche pensabile la messa a punto di una procedura di valutazione che la consideri come una questione a sé, valutando la qualità della ricerca *e* la sua internazionalità. L'assegnazione delle classi di merito, cioè, potrebbe avvenire distintamente per ciascuno di questi criteri, e le fasi successive della valutazione potrebbero considerarli singolarmente. Cioè, mentre nella VQR 2004-2010 l'internazionalità della produzione scientifica era inglobata nel punteggio di qualità della ricerca, in un futuro esercizio potrebbero essere riportati punteggi distinti, permettendo una normalizzazione più efficace degli indicatori a livello di Area e producendo una base evidenziale di maggiore interesse per l'implementazione di politiche mirate.

(Anvur, 2011a, p. 7), dunque queste due caratteristiche supplementari sembrano riferibili essenzialmente al criterio della rilevanza[228]. La distinzione tra i due criteri risulta estremamente simile a quella definita per il REF in cui la rilevanza è intesa «in termini di sviluppo dell'agenda intellettuale del campo» e l'originalità «nei termini del carattere innovativo del prodotto della ricerca» (REF 01.2012, pp. 66-67[229]).

Adottando questa interpretazione dei criteri l'ambiguità si riduce sensibilmente: un prodotto di ricerca è rilevante se costituisce un contributo positivo (nelle diverse sfumature di significato: valido, utile, condivisibile) al dibattito scientifico, originale se presenta argomentazioni o evidenze completamente nuove per questo dibattito.

Alla luce di questa chiarificazione è evidente come la caratteristica essenziale di un prodotto di qualità sia la rilevanza: se un prodotto di ricerca non fornisce un contributo positivo al dibattito scientifico, se non è rigoroso dal punto di vista metodologico o importante dal punto di vista interpretativo non porta con sé alcun valore aggiunto, non può essere considerato di qualità.

Resta da risolvere la questione della co-implicazione apparente tra rilevanza e originalità: da un lato, solo se un prodotto è almeno in parte originale può essere rilevante, dall'altro, solo se il prodotto è rilevante la sua originalità può essere considerata come un indicatore di qualità. Innanzitutto va chiarito che il criterio di originalità non può riferirsi all'ovvia caratteristica dei prodotti della ricerca di costituire contributi inediti al dibattito scientifico. Ciò è evidenziato anche dalla penalizzazione dei casi di plagio (Anvur, 2011)[230], esclusi dalla valutazione proprio in quanto non presentano sufficiente originalità da connotarsi come prodotti della ricerca. Escludendo dunque l'inedicità dei prodotti è necessario riferirsi a un'originalità più connessa ai contenuti dei prodotti (nuove tematiche di studio o prospettive d'analisi, nuovi strumenti di natura tecnica o concettuale) e all'innovatività rispetto agli elementi già presenti nel dibattito (sia in relazione alle interpretazioni che in relazione agli strumenti). La stessa scelta di assegnare al secondo criterio la doppia etichetta di originalità/innovatività lascia propendere per una specificazione del significato che vada nella direzione appena esposta. In questa accezione l'originalità costituisce una effettiva spinta in avanti della conoscenza nel campo solo se accompagnata al rigore metodologico e allo spessore interpretativo che definiscono la rilevanza.

Assumendo dunque la rilevanza come una sorta di criterio essenziale per la valutazione della qualità scientifica di un prodotto, in assenza del quale né l'originalità del lavoro né la sua posizione nel panorama internazionale costituiscono necessariamente indicatori di qualità, originalità e internazionalizzazione si configurano come caratteristiche accessorie della qualità della ricerca.

In altri termini mentre la mancanza di rilevanza di un prodotto ridurrebbe la forza del legame semantico altrimenti esistente tra l'originalità o l'internazionalità e la qualità della ricerca, un basso livello di originalità o internazionalità influirebbe poco o nulla in base alla sua rilevanza scientifica. Si assume dunque che originalità e internazionalità influiscano meno della rilevanza sulla qualità scientifica, ma è doveroso un ulteriore chiarimento: la rispondenza a questi due criteri deve influire

[228] Vale la pena di notare che, considerando le definizioni in uso nell'esercizio inglese (*cfr*. § 3.1) e alcune osservazioni rilevate nel corso delle interviste (nota 220, p. 171), risalta l'assenza, tra i criteri utilizzati per la VQR 2004-2010, del *rigore scientifico*. E' evidente la sostenibilità di un'ipotesi per cui sia il rigore a determinare la rilevanza di un prodotto, nondimeno sembrerebbe auspicabile l'aggiunta di questo criterio a quelli in uso per la valutazione della qualità della ricerca.

[229] Traduzione dall'originale in lingua inglese.

[230] Al punto F: «nei casi accertati di plagio o frode la pubblicazione viene pesata con peso-2»(Anvur, 2011, p. 7).

sull'assegnazione delle classi di merito, altrimenti la procedura non terrebbe conto delle intenzioni del decisore che ha espressamente indicato tre criteri di valutazione. Non è possibile ridurre la qualità scientifica alla sola rilevanza, ma è necessario attribuire a questo criterio un peso differente dagli altri.

Una procedura di sintesi che tenga conto di questa struttura concettuale dovrebbe dunque assegnare al punteggio attribuito al prodotto sulla rilevanza un peso maggiore rispetto al peso dei punteggi relativi a originalità e internazionalizzazione, anche se non è semplice individuare un parametro per determinare quantitativamente i pesi da assegnare ai criteri.

Sarebbe preferibile, per evitare il rischio di una eccessiva discrezionalità, coinvolgere dei "giudici" al fine di stabilire il peso di ciascun criterio nella determinazione della qualità della ricerca di un prodotto (sull'utilità dei giudici nelle procedure di sintesi dei dati si veda: Nobile, 2008)[231]. I giudici, in una procedura simile a quella utilizzata per la selezione degli item nelle scale Thurstone, avrebbero il compito di indicare l'importanza di ciascun criterio ai fini della classificazione di un prodotto dal punto di vista della qualità scientifica. Si tratta di una procedura nota e utilizzabile anche per la valutazione della validità di contenuto (Goode e Hatt, 1952).

Chiaramente in questo caso i giudici dovrebbero essere selezionati all'interno della comunità scientifica di riferimento, se possibile tra i soggetti con maggiore esperienza di valutazione di progetti e/o revisione tra pari. La soluzione più semplice sarebbe utilizzare gli stessi Esperti Valutatori come giudici, anche se la numerosità prevista per l'Area 14, estremamente ridotta[232], farebbe propendere per una più ampia selezione, che coinvolga, ad esempio, anche rappresentanti degli enti da valutare, delle comunità scientifiche, delle maggiori riviste del settore. Rilevare l'opinione di soggetti esperti sull'importanza dei criteri nella determinazione della qualità di un prodotto eliminerebbe la discrezionalità della scelta dei pesi, ma anche in relazione a questa procedura sarebbe auspicabile una propedeutica operazione di ridefinizione dei criteri, che ne specifichi il significato, per evitare ambiguità che complichino la valutazione da parte dei giudici. Una volta rilevata l'importanza dei criteri la procedura di sintesi potrebbe assegnare a ciascuno di essi un peso proporzionale all'importanza rilevata, senza che il decisore stabilisca *ex ante* ed arbitrariamente un peso ai criteri.

La procedura proposta presenta notevoli costi in termini economici e, soprattutto, temporali, nonché una serie di complicazioni connesse agli scopi della valutazione ed agli eventuali conflitti di interesse dei giudici, nonostante i suoi vantaggi una sua applicazione risulta dunque improbabile.

Sebbene la determinazione arbitraria dei pesi non sia argomentabile in base a premesse teorico-concettuali può essere argomentata, almeno in parte, sulla base degli obiettivi della classificazione. Pur in assenza di giudici, sulla base delle considerazioni esposte poco sopra, proponiamo dunque una procedura che assegni un peso maggiore (pari a tre quinti del totale) al criterio della rilevanza:

$$P = 3R + O + I.$$

[231] Sembra vada in questa direzione una osservazione del professor Bonaccorsi circa la possibilità di effettuare pre test delle schede: «per avere anche una base anche di calibrazione che non potrà mai essere perfetta, perché parliamo di numeri anche molto grandi, però che possa migliorare la calibrazione dei giudizi» (Intervista Bonaccorsi).

[232] Per quest'Area sono previsti solo 13 EV, è la penultima Area per ampiezza del GEV dopo l'Area 4 di Scienze della terra (Anvur, 2011, p. 2, Tab. 2).

Il punteggio complessivo per ciascun prodotto assumerebbe valori inclusi tra 5 e 50, di conseguenza le soglie per la determinazione delle classi di merito potrebbero corrispondere a quelle presentate in Tabella 63, adottando l'arrotondamento per eccesso come era per la VQR 2004-2010.

Tabella 63 - Classi di merito per i punteggi calcolati con la formula P=3R+O+I

Classe	Punteggi	Quote corrispondenti
Eccellente	42-50	20,0%
Buono	33-41	20,0%
Accettabile	28-32	11,1%
Limitato	5-27	48,9%

La scelta di pesare la rilevanza per tre è abbastanza argomentabile a partire dai risultati classificatori (Tabella 64), anche se le soglie non sono semplicemente traducibili in asserti sintetici, proprio in ragione della logica matematica e non tipologica sottesa alla procedura. Si noti che originalità e internazionalità contribuiscono a determinare la classe di merito tanto in positivo quanto in negativo, tuttavia rispetto alla procedura 2004-2010:

- un prodotto è *eccellente* non solo se ottiene il massimo punteggio su almeno due criteri, ma anche, ad esempio, se ottiene un punteggio 8 sulla rilevanza e il massimo su originalità e internazionalizzazione, oppure un punteggio 10 sulla rilevanza e la sufficienza (cioè un minimo di 12 punti) sugli altri due criteri;
- un prodotto è *limitato* se ottiene un punteggio inferiore a 3 sulla rilevanza, pur ottenendo il massimo sugli altri due criteri, e un prodotto che risulti eccellente sulla rilevanza (con punteggi 9-10) non può essere classificato come limitato pur ottenendo il minimo su originalità e internazionalità (mentre nella VQR 2004-2010 per rientrare nella classe limitato era sufficiente il punteggio minimo su almeno due criteri, oppure due punteggi mediani e uno minimo).

Tabella 64 – Punteggi (P) e classi di merito (rosso per L, giallo per A, verde chiaro per B e verde scuro per E) per tutti i valori di R per minimo e massimo di O e I

R	R*3	Minimo su O e I			Massimo su O e I		
		O	I	P	O	I	P
1	3	1	1	5	10	10	23
2	6	1	1	8	10	10	26
3	9	1	1	11	10	10	29
4	12	1	1	14	10	10	32
5	15	1	1	17	10	10	35
6	18	1	1	20	10	10	38
7	21	1	1	23	10	10	41
8	24	1	1	26	10	10	44
9	27	1	1	29	10	10	47
10	30	1	1	32	10	10	50

Fino ad ora il filo del ragionamento ha seguito la prima scheda di rilevazione proposta, nella versione semplificata (Riquadro 3), in cui al revisore viene proposta una sola scala di valutazione per ciascun criterio. Nella seconda versione della scheda (Riquadro 4), invece, a ciascun criterio corrispondono due o tre punteggi, ciascuno riferito a un diverso referente. E' evidente che «quanto più segmentata è la procedura di classificazione, tanto più diventa complicata la ricombinazione dei giudizi parziali» (Agodi, 1999, p. 130).

Qui i punteggi rilevati sono riferiti a uno stesso concetto (il criterio) su diversi referenti, non si tratta di diversi indicatori di uno stesso concetto ma di uno stesso indicatore riferito a diversi oggetti. La questione su cui riflettere prima di proporre una procedura di sintesi è solo apparentemente semplice: le proprietà da rilevare devono essere distribuite su tutti i referenti oppure possono riferirsi a uno solo di essi? In altri termini, un prodotto per essere considerato rilevante o originale deve esserlo contemporaneamente sul piano teorico, metodologico e sostantivo oppure è sufficiente che lo sia su uno solo di questi piani?

Dato che le definizioni non possono offrire soluzioni a questo problema, e che dal punto di vista logico e semantico entrambe le opzioni possono apparire altrettanto sostenibili, è opportuno considerare il punto di vista pragmatico. Non tutti i prodotti della ricerca sono valutabili con uguale rilievo in riferimento a tutti e tre i referenti scelti per i primi due criteri (rilevanza e originalità); è cioè possibile, ad esempio, che un contributo sia di natura quasi esclusivamente teorico-concettuale e in questo caso la rilevanza sugli altri referenti sarebbe pressoché nulla. Considerando ad esempio un prodotto estremamente rilevante per il suo approccio teorico-concettuale (10), ma riferito a un problema d'indagine non particolarmente rilevante (6) e senza riferimenti a specifici approcci metodologici (1), non è sostenibile che la valutazione complessiva corrisponda alla somma (17) o alla media (5,6) dei punteggi ottenuti per i tre referenti: la mancanza di rilevanza con riferimento al problema o alla tecnica non può sminuire la rilevanza teorica del contributo. Lo stesso ragionamento può essere esteso a contributi centrati su specifiche problematiche o su questioni di natura prettamente metodologica.

Assumere che la rilevanza di un prodotto corrisponda alla sua massima rilevanza (indipendentemente dal referente cui questa è riferita) rende invece più sostenibile anche il confronto diretto tra prodotti della ricerca in relazione ai quali il criterio sia riferibile a un numero diverso di referenti. La scelta pragmaticamente più corretta sembra, dunque, quella di considerare per ciascun prodotto solo il punteggio massimo, per cui nell'esempio (dove a=10, b=6 e c=1) il prodotto otterrebbe 10 sulla rilevanza. Una volta determinati i punteggi massimi su ciascun criterio la sintesi potrebbe procedere esattamente come per la scheda sintetica, tramite una somma ponderata dei punteggi e la riconduzione in classi tramite soglie.

Una procedura di sintesi di questo tipo, anche se basata su pesi definiti da giudici qualificati, risolverebbe però solo in parte il problema relativo all'ambiguità presente nella definizione delle classi di merito circa il criterio relativo all'internazionalizzazione, e risulterebbe coerente solo in parte con la definizione delle classi in relazione all'originalità, che invece viene esplicitamente citata come caratteristica tanto della classe *eccellente* quanto della classe *buono*. Il rispetto di questi presupposti rende necessario un approccio tipologico e non matematico, a partire da classi di merito assegnate sui singoli criteri. Definendo le classi in base al criterio quantitativo nel caso di schede con scale 1-10 la classe *eccellente* corrisponderebbe ai punteggi 9 e 10, *buono* ai punteggi 7 e 8, *accettabile* al punteggio 6 e *limitato* ai punteggi da 1 a 5.

La tipologia che si ottiene rispettando alla lettera le definizioni semantiche delle classi di merito (Anvur, 2011) lascia però fortemente a desiderare, dato che come già sottolineato (*cfr*. § 3.2) la definizione delle classi non è esaustiva. Non considerando tutte le possibili combinazioni tra i criteri, le definizioni non consentono di classificare tutti i tipi risultanti da uno spazio degli attributi che combini rilevanza, originalità e internazionalità (Tabella 65), ma soltanto parte di essi. Le problematiche evidenziate circa la *boundary indefiniteness*, la *membership indefiniteness* e la *cut-off*

indefiniteness dei criteri e il loro riflesso sulla definizione delle classi rendono pressoché impossibile stabilire chiaramente le regole di classificazione.

Nuovamente risulterebbe necessaria una specificazione delle definizioni, proprio perché propedeutica alla progettazione di procedure di sintesi metodologicamente appropriate.

Tabella 65 – Classificazione tipologica della qualità della ricerca in base a rilevanza, originalità e internazionalizzazione in base alle definizioni delle classi di merito

Rilevanza	*Originalità*	*Internazionalizzazione*			
		L	A	B	E
L	L	L	L	L	
	A				
	B				
	E				
A	L				
	A	A	A	A	A
	B				
	E				
B	L				
	A				
	B	B	B	B	B
	E				
E	L				
	A				
	B				
	E	E	E	E	E

E' possibile immaginare diversi modi di classificare i tipi non contemplati nelle definizioni delle classi di merito, qui si presenta una sola soluzione ipotetica che rispetti sia i riferimenti disponibili, sia l'assunto che la qualità scientifica di un prodotto sia legata innanzitutto dalla sua rilevanza, poi a originalità e internazionalità (Tabella 66).

Tabella 66 - Classificazione tipologica della qualità della ricerca in base a rilevanza, originalità e internazionalizzazione estesa a tutti i tipi a partire dalle definizioni delle classi di merito

Rilevanza	*Originalità*	*Internazionalizzazione*			
		L	A	B	E
L	L	L	L	L	A
	A	L	L	L	A
	B	L	L	L	A
	E	A	A	A	A
A	L	L	A	A	A
	A	A	A	A	A
	B	A	A	A	B
	E	A	A	B	B
B	L	A	A	B	B
	A	A	B	B	B
	B	B	B	B	B
	E	B	B	B	E
E	L	B	B	B	E
	A	B	B	E	E
	B	B	E	E	E
	E	E	E	E	E

Rispetto alla procedura di sintesi matematica in questa soluzione tipologica il numero di casi in cui la valutazione ottenuta sui criteri di originalità e internazionalità abbassa la valutazione basata sulla rilevanza è molto ridotta, tuttavia in mancanza di una struttura concettuale chiara resta complesso giustificare adeguatamente le scelte operate circa la classificazione finale dei prodotti.

Fino a questo punto si è tentato di proporre alcuni possibili aggiustamenti alle procedure utilizzate per la VQR, senza mettere in discussione le definizioni dei criteri di valutazione. Si è tentato, attraverso la riflessione sulle possibilità di specificazione del significato e delle relazioni tra i criteri, di progettare delle definizioni operative che risultassero (*ex ante*) il più affidabili possibile. I risultati ottenuti, però, sono solo in parte soddisfacenti. L'ambiguità e la vaghezza delle definizioni rendono sostanzialmente impossibile la messa a punto di una definizione operativa che risulti metodologicamente adeguata e pienamente argomentabile, spingendo a richiamare l'opportunità di una ri-definizione, da parte del decisore (cioè del Ministero) e dell'Agenzia, dei criteri di valutazione e delle classi di merito.

Al di là degli obiettivi strategici della valutazione l'individuazione dei criteri dovrebbe tenere conto soprattutto del parere della comunità scientifica. La definizione della qualità della ricerca è al di là degli obiettivi di questo lavoro. Il dibattito sulla VQR ha incluso diversi interventi riferiti direttamente al contenuto dei criteri (ad esempio, La Rocca, 2013), ciò nonostante non sono rinvenibili proposte strutturate e compiute in riferimento alla loro definizione. Un reale coinvolgimento delle comunità nel processo decisionale costituirebbe un investimento notevole di tempo e risorse, un investimento a lungo termine, e andrebbe riferita alle singole discipline.

Dando per scontata la possibilità di una consultazione diretta con le strutture di ricerca e le società scientifiche, si intende qui proporre la realizzazione di studi finalizzati all'individuazione di criteri valutativi che risultino il più condividivisi possibile dai particolari punti di vista delle comunità scientifiche che su quella base dovranno essere valutate. Un esempio è rappresentato dallo studio di Becker e colleghi (2006) mirato alla definizione della qualità nella ricerca scientifico-sociale, in base a quanto emerso dalla rilevazione delle esperienze e delle opinioni dei ricercatori e degli altri *stakeholders*, in vista della revisione delle procedure di valutazione che nel Regno Unito hanno condotto alla messa a punto del REF (McNay, 2003; *cfr*. § 3.1).

Una possibilità simile, ma ancora più centrata sulla necessità della costruzione di un consenso più ampio intorno ai criteri della valutazione è individuabile nell'approccio proposto da Hug e Ochser (2014), in cui a partire da una indagine esplorativa, utile alla definizione dei criteri in base a ciò che gli appartenenti a una comunità definiscono come qualità, validi questi criteri in indagini più estese, contribuendo a costruire un consenso attorno alle definizioni e agli indicatori così elaborati. Con particolare riferimento alle scienze umane e sociali gli autori sottolineano che «dal momento che le nozioni di qualità dei ricercatori, spesso esistono come conoscenza tacita (cioè, come una conoscenza che non può essere articolata in modo semplice e chiaro, Polanyi, 1967), sono necessari metodi che traducano le tacite di qualità in conoscenza esplicita (cioè, conoscenza che può essere espressa chiaramente)» (*ivi*, p. 63).

Tornando alla procedura di rilevazione dei giudizi, una modifica estremamente semplice e poco costosa da implementare è l'eliminazione della possibilità per i revisori di rivedere la classe di merito una volta assegnati i punteggi. In relazione alla opportunità di una correzione degli eventuali errori materiali sarebbe, infatti, possibile richiedere conferma dei punteggi assegnati sui singoli

criteri, mentre la formazione dei revisori e l'esposizione chiara dei criteri di classificazione eliminerebbero lo scopo di auto-formazione segnalato dall'Anvur in relazione a questa possibilità.

L'ultimo passaggio della sintesi è la riconduzione delle classi assegnate dai singoli *referee* alla classe di merito finale. Nella procedura di valutazione dei prodotti nella VQR questa fase avveniva nell'ambito dei gruppi di consenso appositamente costituiti in seno ai GEV. Nel caso in cui il prodotto avesse ottenuto valutazioni concordanti o discordanti di una sola classe da parte dei due revisori il gruppo di consenso aveva il compito di approvare la classe di merito finale stabilita tramite il confronto della somma dei punteggi assegnati al prodotto dai due revisori con delle soglie teoriche (Anvur, 2014; *cfr.* § 4.3.2). Nel casoin cui, invece, il prodotto avesse ottenuto valutazioni discordanti di più di una classe il gruppo di consenso aveva il compito di selezionare un terzo revisore e in seguito riprendere visione e ridiscutere la classificazione del prodotto.

In questo passaggio la discussione interna ai gruppi di consenso è centrale: solo leggendo le valutazioni dei *referee,* e discutendone, il gruppo può pervenire a una valutazione finale condivisa. Nessuna procedura automatizzata che preveda un confronto tra classi o punteggi può infatti identificare eventuali *bias* legati ai punti di vista disciplinari o paradigmatici dei valutatori e considerarne gli effetti sull'esito della classificazione, mentre indubbiamente una attenta analisi da parte di un gruppi di esperti potrebbe non solo identificare ma anche correggere (direttamente o tramite l'assegnazione del prodotto a un terzo revisore) questo genere di distorsioni.

In quest'ottica il fatto che le valutazioni siano concordanti, cioè che il prodotto abbia ottenuto la stessa classe di merito da entrambi i revisori, non vuol dire che queste non debbano essere discusse nell'ambito dei gruppi di consenso. Se la responsabilità della valutazione fa capo al GEV è il GEV a dover visionare, discutere ed approvare le valutazioni dei prodotti.

In questa fase la compilazione da parte di tutti i revisori del campo dedicato alla motivazione dei giudizi e all'annotazione di notizie rilevanti per gli EV sarebbe di grande utilità ai fini della discussione e della identificazione della classe di merito finale nell'ambito dei gruppi di consenso. Lo stesso GEV 14 ha evidenziato l'utilità dei commenti dei revisori, suggerendo di rendere obbligatorio il campo: «la presenza delle motivazioni, infatti, si è rivelata uno strumento molto utile per la convalida definitiva della valutazioni nella discussione interna ai *consensus groups*, mentre la sua assenza ha reso necessario in molti casi il ricorrere ad un terzo referaggio» (Anvur 2013d, GEV 14, p. 65)[233].

Una attenta progettazione della scheda di rilevazione, il suo pre test, la messa a punto di procedure di sintesi adeguate sono in strettissima relazione con la qualità dei dati, cioè con la validità e l'attendibilità delle valutazioni. In relazione alla pubblicità, alla controllabilità e alla replicabilità della procedura è fondamentale non solo rendere pubbliche le schede di rilevazione e le procedure di sintesi, ma anche argomentare adeguatamente tutte le scelte effettuate nel corso della progettazione dell'esercizio.

6.1.2 Proposte per la valutazione bibliometrica

La definizione operativa della qualità della ricerca utilizzata per la valutazione diretta tramite analisi bibliometrica presenta problematiche molto diverse rispetto alla valutazione in peer review,

[233] E' nelle intenzioni dell'Anvur rendere obbligatorio il campo: «intendiamo anche rendere obbligatoria la scrittura di un commento alle valutazioni» (Intervista Benedetto).

ma anche in questo caso le proposte che possono essere avanzate non costituiscono soluzioni definitive, ma possono rappresentare passi in avanti verso una procedura di valutazione dei prodotti più aderente ai criteri e meno esposta a rischi di distorsione, pur non essendone esente.

E' chiaro che la maggior parte delle problematiche tecniche derivano dalla scelta di utilizzare più di un database, ma prima di discutere questo aspetto sono messe a fuoco le questioni relative alla selezione degli indicatori, alla loro riconduzione alle classi di merito e alla sintesi finale, assumendo che i database utilizzati siano due: WoS e Scopus. Inoltre, è importante sottolineare che gran parte delle possibili distorsioni sono legate alla natura bibliometrica dei dati e alle modalità di costruzione e gestione dei database e che dunque non sono eliminabili a meno di controllare e correggere pesantemente le basi di dati disponibili.

La selezione degli indicatori pone delle questioni sia in relazione alla loro aderenza con la formulazione dei criteri sia in relazione alla loro comparabilità, non tanto con riferimento al conteggio delle citazioni, quanto con riferimento agli indicatori di impatto delle riviste: l'*impact factor* e lo SJR. Si tratta di indici molto diversi che, come argomentato nel Capitolo 4 e in parte mostrato con l'analisi presentata nel Capitolo 5, possono condurre a valutazioni differenti di uno stesso prodotto.

Dal punto di vista metodologico sarebbe opportuno scegliere un'unica definizione operativa dell'impatto della rivista e applicarla in entrambi i database. In questo caso sarebbe più semplice adottare l'algoritmo di calcolo dell'*impact factor* anche in Scopus, dato che l'algoritmo di calcolo dello SJR è più complesso e non sembra plausibile una sua applicazione all'interno di WoS.

Nel panorama degli indici bibliometrici non esistono indicatori in grado di cogliere aspetti specifici della qualità scientifica, come l'originalità, la rilevanza o l'internazionalizzazione. E' relativamente semplice immaginare una misura dell'internazionalità (ad esempio una sorta di H-index riferito ai singoli articoli che tenga conto della provenienza delle citazioni per nazionalità in cui l'indice assuma valore *n* pari al numero di nazioni da cui si ricevono almeno *n* citazioni potrebbe cogliere l'ampiezza dell'impatto citazionale internazionale), ma l'operazione risulta molto più complessa, se non impossibile per la rilevanza e l'originalità. Inoltre l'effettiva validità e attendibilità di indicatori di questo genere andrebbe discussa lungamente nella comunità scientifica prima che sia possibile pensare a un loro utilizzo negli studi scientometrici, e una fase di validazione ancora più estesa perché siano utilizzabili in esercizi di valutazione della ricerca come la VQR.

Una misura più sofisticata dell'impatto, in grado di tenere in considerazione un maggior numero di problematiche e di assicurare una maggiore comparabilità, potrebbe ad esempio essere rappresentata dallo SNIP (*source normalized impact per paper*), proposto da Moed (2010). Quest'indice considera al numeratore il numero di citazioni nell'anno in analisi sul numero di articoli pubblicati nei tre anni precedenti[234], e al denominatore il rapporto tra il potenziale citazionale nel database[235] di ciascun campo disciplinare[236] e il potenziale citazionale della rivista mediana nel

[234] Il numeratore è dunque simile all'impact factor, ma con una differente finestra temporale, inoltre non conteggia le citazioni di tipi di documenti non sottoposti a peer review (articoli, conference papers e reviews), limitando l'influenza delle auto-citazioni editoriali e le citazioni di documenti "non citabili" come lettere e note (Moed, 2010).

[235] Il potenziale citazionale nel database per le categorie tematiche è calcolato come la media delle citazioni di 1-3 anni per ciascuna rivista e pubblicati in riviste indicizzate nel database (Moed, 2010).

[236] Operativizzato non utilizzando le categorie tematiche ma sulla base delle citazioni tra gli articoli: il campo di studio qui è inteso come l'insieme degli articoli che citano un dato documento in un dato lasso di tempo. Si

database. Si tratta di un indicatore di impatto citazionale "*contestuale"*, che considera le caratteristiche del campo disciplinare come ad esempio la frequenza con cui gli autori citano altri papers, la rapidità di maturazione dell'impatto citazionale e il grado con cui un database copre la letteratura del campo. I vantaggi posti da quest'indice sono numerosi, ma restano anche diverse criticità, come la sovrastima dell'impatto delle riviste che pubblicano numerose review, l'inaffidabilità delle classificazioni dei documenti, la crescita progressiva della letteratura in un dato campo disciplinare (Moed, 2010). L'indice, implementato tra gli altri in Scopus, è stato discusso in letteratura e varie modifiche sono state proposte (per tutti, Waltman *et al.* 2013), ma alcune problematiche sono chiaramente non risolvibili con la normalizzazione. Nessuna di queste soluzioni eliminerebbe i dubbi circa l'utilizzo di una misura di impatto della rivista per la valutazione degli articoli in essa contenuti, né sembra possibile immaginare soluzioni che non tengano conto di un indicatore di questo genere nell'ambito di un esercizio come la VQR. Dato infatti il breve lasso di tempo che intercorre tra il periodo di riferimento della valutazione e la valutazione stessa sembra impossibile basarsi esclusivamente sul numero di citazioni ricevute dal documento, ed è necessario utilizzare altre fonti, ad esempio l'impatto della rivista considerato come un indicatore predittivo dell'impatto degli articoli in essa contenuti.

La procedura di riconduzione dei singoli indicatori alle classi di merito, per quanto adeguata al controllo delle distorsioni macro che possono occorrere nel confronto tra indici bibliometrici, è sicuramente ancora perfettibile, così come la sintesi finale tramite le matrici quadrate[237].

Innanzitutto, come già segnalato, sarebbe opportuna una maggiore attenzione alla comparabilità delle finestre temporali per il conteggio delle citazioni (*cfr*. § 4.2.3). Nella VQR, infatti, la finestra temporale per gli articoli pubblicati nel 2004 è di sette anni, per quelli pubblicati nel 2005 di sei, per quelli pubblicati nel 2006 di cinque, e così via fino a giungere a una finestra temporale di un solo anno per le pubblicazioni del 2010.

Sarebbe più corretto utilizzare una finestra temporale identica per tutti gli anni di pubblicazione, indipendentemente dall'ampiezza della finestra temporale su cui i dati risultano disponibili. Nel caso dell'Area 3, che ha utilizzato due matrici di classificazione differenti in base all'anno di pubblicazione (2004-2008 o 2009-2010, *cfr.* § 4.3.3), modificando al minimo la procedura, dunque tenendo fermi i due periodi e volendo impiegare in entrambi i casi la finestra temporale più ampia possibile, una soluzione semplice sarebbe stata quella di utilizzare una finestra di tre anni per i

noti che i campi sono delimitati su misura e possono includere riviste generali o multidisciplinari (Moed, 2010).

[237] Il Presidente Anvur ha sintetizzato così le possibili modifiche: «per quanto riguarda la bibliometria miglioreremo questa matrice, cioè l'interdipendenza di questi indicatori, sappiamo già come migliorarla; forse entreremo un pochino nel dettaglio delle classifiche, se agli eccellenti dare un peso piuttosto che un altro, oppure considerarli sul 10-15% perché… questo è stato argomento di discussione, ma insomma in buona sostanza non andremo a cambiare molto» (Intervista Fantoni); e il Presidente del GEV 3 ha evidenziato i margini di perfettibilità delle procedure: «io vorrei rivedere un po' i casi in cui non c'è concordanza fra i due aspetti, cioè la qualità della rivista e il numero di citazioni. Sono abbastanza convinto che questi due criteri da vedere… ripeto più classe di merito che impact factor per la verità, e il numero di citazioni, sono due criteri corretti il cui peso relativo deve dipendere dagli anni, però questo va un attimo rivisto. Adesso abbiamo l'esperienza della VQR, quindi delle simulazioni sensate si possono fare questa volta e questo fa la differenza» (Intervista Barone).

prodotti pubblicati tra il 2004 e il 2008[238], e una finestra di un solo anno per i prodotti pubblicati nel 2009 o nel 2010[239]. Questo accorgimento avrebbe reso perfettamente comparabili tra loro le classificazioni destinate a contribuire all'assegnazione della classe di merito finale tramite ciascuna matrice.

In secondo luogo è necessario riflettere sull'uso delle *subject categories* e delle classi dell'ASJC. Si è già sottolineato che queste classificazioni tematiche sono riferite alle riviste e non ai singoli articoli in esse contenute e dunque non sono in grado di assicurare che la procedura confronti gli articoli all'interno di campi disciplinari omogenei (*cfr*. § 4.2.3). Lo scopo della VQR è però la valutazione di singoli prodotti, dunque sarebbe preferibile utilizzare una classificazione che abbia come oggetto gli articoli, non le riviste. Una possibile soluzione sarebbe prevedere la classificazione dei singoli articoli per campi disciplinari omogenei sfruttando le citazioni stesse, come per la definizione dei campi di studio usate per il calcolo dello SNIP (Moed, 2010). Le possibili procedure sono diverse e sono disponibili diverse proposte, si veda ad esempio quella di Waltman e van Eck (2012), che origina una classificazione gerarchica e che dunque permetterebbe di fare riferimento a categorie più o meno ampie a seconda degli obiettivi. Queste procedure di classificazione, tuttavia, richiederebbero un enorme investimento di tempo e risorse. E' la necessità di un grande investimento iniziale a fronte dell'assenza di assicurazioni sulla effettiva efficacia e validità delle classificazioni ottenute a giustificare, in una prospettiva economica e pragmatica, l'utilizzo di categorie pre-costituite, come è stato nella VQR.

La problematica centrale e contemporaneamente la meno semplice da risolvere è indubbiamente la calibratura degli algoritmi. La procedura in uso infatti riconduce i singoli indicatori, cioè il numero di citazione e l'indice di impatto delle riviste, alle classi di merito facendo riferimento alla definizione quantitativa delle classi: 20% eccellente, 20% buono, 10% accettabile e 50% limitato[240], tuttavia queste quote dovrebbero corrispondere all'esito della classificazione e non ai passaggi intermedi (*cfr*. § 4.3.3).

[238] Conteggiando, ad esempio, per le pubblicazioni del 2004 le citazioni ricevute fino al 31/12/2007, per i prodotti del 2005 le citazioni ricevute fino al 31/12/2008, e così via.

[239] Conteggiando, cioè, per le pubblicazioni del 2009 le citazioni fino al 31/12/2010 e per le pubblicazioni del 2010 le citazioni fino al 31/12/2011.

[240] Gran parte dei suggerimenti di modifica alla procedura bibliometrica, sia da parte degli EV che da parte dei membri del Consiglio direttivo Anvur è riferita alla scelta di queste percentuali, che pure va oltre le prerogative dell'Agenzia, ad esempio: «cambiare le soglie. Quelle probabilmente sono state abbastanza sbagliate. Cioè il 20% di eccellenza è troppo. Un'eccellenza potrebbe essere probabilmente intorno al 7-8%, io cambierei. Oppure uno decide che mette, come per l'ERC, l'out-standing, poi le eccellenze, e poi.. però dividere, mettere una categoria che sia relativa a qualche percento top io la farei. Tanti prodotti sono finiti in categorie alte, anche questo bisogna dire. Per cui probabilmente trovare un meccanismo che permetta di scandagliare meglio la zona grigia dei prodotti non particolarmente buoni» (Intervista Pacchioni); da rivedere sarebbe l'ampiezza delle classi, non il numero, dato che le quattro classi: « permettono di ragionare su elementi che somigliano ai quartili, anche se in effetti per l'eccellenza il quartile è un po' troppo grande. E' quindi meglio ragionare con il top 10 o il top 20%, che sono altre varianti possibili del gioco. Forse lo standard internazionale vorrebbe un top 10% poi 20 % poi un ulteriore 30% e il restante 40%, se dovessi dare una scala ideale. Perché la fascia bassa deve essere più ampia, per evidenti ragioni, e mano a mano che ti avvicini all'alto devi rastremare la metrica, altrimenti hai un effetto di sovra rappresentazione. Nei confronti internazionali quello che si usa è il top 10% o addirittura il top 1%, però il top 1% è un caso che poi farebbe poche unità, non rileva… però il 10 è spesso usato. Se uno volesse dire l'eccellenza italiana confrontata con l'eccellenza europea, l'ERC lavora sul 10 % la National Science Foundation lo stesso […] Da questo punto di

La riconduzione degli indicatori bibliometrici (considerati singolarmente) alle classi di merito e la sintesi per via tipologica non conducono necessariamente a una classificazione che rispetti le quote previste dal bando[241]. E' chiaro infatti che le distribuzioni dei due indicatori, gli eventuali pareggi e gli esiti della combinazione delle due classificazioni possono influire sensibilmente sulle quote. In base ai risultati della VQR, in cui le distribuzioni ottenute per l'universo dei prodotti non corrispondono alla distribuzione teorica prevista per le classi di merito, la necessità di modificare le soglie al fine di ricalibrare l'algoritmo è più che evidente (*cfr*. § 4.3.3) ed è stata esposta dall'Anvur stessa in appendice al rapporto finale (Anvur, 2011a, Appendice A).

La calibratura delle procedure non è un obiettivo semplice. Infatti, volendo ottenere una classificazione finale dell'universo dei prodotti (cioè dell'insieme degli articoli indicizzati e non solo dei prodotti sottomessi a valutazione) che rispetti le quote delle classi di merito, è necessario tenere conto non solo della probabilità che un prodotto ottenga la classe X su ciascun indicatore, ma anche della combinazione di queste probabilità e dell'esito della valutazione peer dei prodotti classificati come *undecided* dalla procedura bibliometrica. La categoria tematica, il tipo di documento, l'anno di pubblicazione sarebbero naturalmente solo alcune delle variabili da considerare nella messa a punto del sistema di calibrazione, senza considerare, in quasi tutte le Aree, l'uso parallelo di due diversi database. A questo proposito va sottolineato che la simulazione dell'Anvur per la calibrazione degli algoritmi bibliometrici (*cfr*. § 4.3.3) è stata condotta esclusivamente su WoS e che le problematiche relative alla calibrazione su più database non vengono toccate (Anvur, 2011a, Appendice A).

La disponibilità dei dati relativi alla VQR 2004-2010 può costituire una buona base di partenza per la riflessione e la messa a punto di algoritmi più aderenti alla definizione delle classi, permettendo anche un'analisi approfondita degli esiti di diversi algoritmi utilizzati dai singoli GEV.

6.2 La rilevazione della qualità

Le modalità di rilevazione della qualità della ricerca, in particolare quelle di selezione delle fonti per ciascun prodotto, sono centrali al fine della determinazione della qualità dei dati almeno quanto le definizioni operative. L'analisi presentata nel Capitolo 5 ha messo in luce i punti cruciali tanto con riferimento alla procedura di valutazione tramite peer review quanto con riferimento alla procedura di valutazione diretta tramite analisi bibliometrica.

---

vista sarà difficile evitare che il ministero voglia mettere qualche sua opinione, ne riparleremo tra qualche mese» (Intervista Bonaccorsi).

[241] Nelle intenzioni, per la prossima VQR: «l'affinamento significativo sarà che questa volta avremo il tempo di effettuare una calibrazione ex ante e che quindi tutte le aree bibliometriche avranno assolutamente la stessa calibrazione. Non so se lei ha visto come è stata fatta la calibrazione, ma in sostanza quando si dice che i prodotti eccellenti costituiscono il 20% della produzione scientifica mondiale in una certa area, l'idea era che applicando i criteri bibliometrici del GEV a tutto quello che c'era nel database alla fine venivano classificati come eccellenti il 20% degli articoli contenuti nel database. Questa operazione è stata fatta a livello di area, è stata fatta un po' in fretta, è stato lasciato un margine di libertà eccessivo ai GEV, per cui alla fine la probabilità di ottenere una valutazione eccellente per i vari GEV all'interno del database ISI o Scopus non era la stessa. Questa volta invece sarà la stessa, non solo, ma la calibrazione la faremo anche per *subject category* per cui non solo l'area, ma tutte le SC avranno la stessa calibrazione in tutte le aree scientifiche» (Intervista Benedetto).

La scelta dei revisori e la distribuzione dei prodotti sono in grado di influire sensibilmente sulla rispondenza delle valutazioni rilevate alle condizioni logiche e metodologiche da rispettare perché l'esercizio di valutazione risulti valido. La qualità dei revisori è il primo requisito per una peer review affidabile, il profilo scientifico e l'esperienza dei pari coinvolti sono essenziali, sarebbe dunque opportuno ridurre il rischio di selezionare revisori secondo criteri eterogenei e incostanti nel corso dell'esercizio.

La seconda questione su cui centrare l'attenzione è l'assegnazione dei prodotti ai revisori non solo in relazione alle caratteristiche di autori e revisori oppure alle tematiche trattate, ma anche per questioni più banali ma non meno rilevanti come le differenze nel carico di lavoro, nella varietà dei prodotti per i singoli revisori, oppure nella varietà dei revisori per i prodotti di uno stesso autore. Non andrebbero inoltre sottovalutati i possibili benefici di una breve formazione dei revisori in relazione all'interpretazione dei criteri e delle classi di merito, all'uso delle schede di rilevazione e alle questioni etiche connesse alla peer review.

In relazione alla stabilità e all'uniformità delle scale di giudizio dei revisori, vengono presentati alcuni spunti di ricerca. I dati della VQR 2004-2010 e quelli che potrebbero essere prodotti da un pre test delle schede di rilevazione costituirebbero una base empirica di grande interesse su questi temi, anche se un reale approfondimento, come si avrà modo di argomentare, necessiterebbe della conduzione di studi *ad hoc*.

Con riferimento alla procedura di peer review la rilevazione della qualità è invece completamente affidata ai database bibliometrici. A questo riguardo l'uso di più database rappresenta una questione centrale che meriterebbe un approfondimento e una riflessione più attenta, anche in relazione alla questione della calibratura degli algoritmi. Un maggiore controllo degli esiti della valutazione diretta tramite analisi bibliometrica sarebbe inoltre estremamente auspicabile, sia in ragione del grado di affidabilità *a priori* dei dati bibliometrici (van Raan, 1996) sia al fine di effettuare un controllo della loro affidabilità *a posteriori*[242], prima che gli esiti delle procedure possano costituire una base informativa per decisioni di natura gestionale e/o finanziaria.

6.2.1 Proposte per la valutazione in peer review

La selezione dei revisori è cruciale in tutte le procedure di peer review: solo se il profilo scientifico e l'esperienza dei pari coinvolti sono adeguati al compito le valutazioni fornite saranno attendibili e valide. E' evidente che, a questo proposito, rispetto alla VQR 2004-2010 sarebbero più che opportuni alcuni aggiustamenti. E' il caso di definire con maggiore chiarezza i profili dei revisori; in particolare sarebbe auspicabile una indicazione esplicita dei requisiti *minimi* cui ciascuno di essi deve rispondere per poter partecipare all'esercizio come valutatore dei prodotti.

In primo luogo, va affrontata la questione dell'ambito di *expertise* dei revisori. Nella VQR quest'ambito era delimitato dai settori-scientifico disciplinari[243] e ulteriormente caratterizzato

[242] In senso marradiano si intende per affidabilità a priori il grado di fiducia riposto nella definizione operativa e per affidabilità a posteriori la valutazione da parte del ricercatore del grado di affidabilità dell'esito di una definizione operativa, a seguto della rilevazione e dell'analisi dei dati (Marradi, 1990b).

[243] Agli SSD in alcune Aree si affiancavano obbligatoriamente altre classificazioni, ad esempio la classificazione ERC (*European Research Council*), ma in molte altre queste classificazioni aggiuntive erano opzionali. A

tramite alcune parole chiave segnalate dai revisori stessi nel modulo di auto-candidatura. Si aprono due questioni: i settori scientifico-disciplinari sono spesso categorie macro ed eterogenee al loro interno[244], inoltre il settore di appartenenza del revisore o la sua auto-classificazione potrebbero rispecchiare variamente la sua concreta attività scientifica.

Un modo relativamente semplice per identificare il settore di competenza dei revisori potrebbe basarsi sulla loro produzione scientifica recente. Ad esempio sarebbe possibile chiedere a ciascun candidato a revisore di presentare un certo numero di prodotti (ad esempio 5, ma la questione sarà ripresa più avanti), selezionando i migliori all'interno della propria produzione scientifica più recente (ad esempio degli ultimi 3 anni), e di ricondurre ciascun prodotto a una classificazione tematica elaborata *ad hoc*. Ciascun prodotto potrebbe essere ricondotto a più categorie tematiche, tuttavia, per ciascuno di essi, una delle categorie dovrebbe essere indicata dall'autore come la categoria principale. In sostanza l'esito di questa procedura sarebbe una classificazione dei revisori per categoria tematica, sicchè ciascun revisore potrebbe essere direttamente legato a una o più classi principali e a una o più classi secondarie.

In questo modo sarebbe possibile basare effettivamente l'individuazione dei campi di studio dei revisori sulla loro esperienza di ricerca, sfruttando la classificazione tematica della loro produzione scientifica recente. Il riferimento alle pubblicazioni degli ultimi anni infatti permette di identificare le principali tematiche di ricerca su cui ciascun revisore ha lavorato in tempi recenti, su cui dunque ha un certo grado non solo di conoscenza, ma anche di aggiornamento rispetto alla letteratura e più in generale al dibattito scientifico.

Il passo più difficile è effettivamente la costruzione di una classificazione adeguata allo scopo, che non abbia un livello di generalità troppo elevato, come gli SSD, ma non risulti neppure troppo minuta, creando il rischio di limitare, nelle fasi successive dell'esercizio, la selezione dei revisori per specifici prodotti a micro-comunità scientifiche, con tutti i rischi che ne deriverebbero.

Le declaratorie dei settori scientifico-disciplinari potrebbero costituire lo spunto per una soluzione che vale la pena prendere in considerazione (Allegato B al DM 4 ottobre 2000). Ciascuna definizione infatti fa riferimento a diverse tematiche di indagine e campi di specializzazione che, pur essendo sufficientemente ampi non risultano eccessivamente vaghi né troppo specifici. A titolo di esempio si riportano le definizioni del settore di Sociologia Generale, molto ampio ed eterogeneo, e il settore di Sociologia dei Fenomeni Politici, più specialistico e omogeneo, entrambi inclusi nell'Area delle Scienze Politiche e Sociali:

> **Sociologia Generale (SPS/07)**
> Il settore contiene una serie di campi di competenza concernenti la propedeutica teorica, storica e metodologica della ricerca sociale, i confini epistemologici della sociologia, gli strumenti teorico-metodologici e le tecniche per l'analisi delle processualità micro e macro-sociologiche. In quest'ottica si articola in varie aree che vanno dalla sociologia in generale (per le prospettive teoriche fondamentali, il linguaggio delle scienze sociali, l'ordine e il mutamento

proposito della classificazione ERC si veda il documento: *Guide for Applicants in the ERC Work Programme 2014*, pp. 44-53; http://erc.europa.eu/sites/default/files/document/file/info_for_applicants_adg_2014.pdf.

[244] Curiosamente una indicazione in questo senso è venuta dal Presidente di Area 3: «ci vogliono degli aggiustamenti, allora l'aggiustamento fondamentale che ci vorrebbe è che i settori disciplinare sono sbagliati e infatti […] i PRIN e i FIRB sono fatti sui settori europei che non sono il meglio al mondo ma sono un po' più decenti di come siamo conciati adesso» (Intervista Barone).

e per le categorie e le problematiche relative al rapporto teoria-ricerca empirica), alla metodologia e tecnica della ricerca sociale, alla politica sociale connessa alle diverse tipologie di welfare, ai metodi e alle tecniche del servizio sociale ai sistemi sociali comparati, all'analisi dei gruppi, della salute della scienza, dello sviluppo, della sicurezza sociale, ai metodi della pianificazione, alla storia del pensiero sociologico.

Sociologia dei Fenomeni Politici (SPS/11)
Il settore contiene una serie di campi di competenza concernenti il rapporto fra la società e il mondo delle decisioni strategiche vincolanti, dal parlamento, al governo, ai partiti politici, all'analisi del rapporto sistemi sociali-politiche pubbliche, talvolta anche in una prospettiva internazionalistica, dall'analisi socio-politica in generale allo studio sociologico dell'amministrazione, alla sociologia delle relazioni internazionali, alla comunicazione politica.

A partire dai contenuti di queste declaratorie sono individuabili, più o meno chiaramente, una serie di categorie tematiche "meso":

Sociologia Generale (SPS/07)

1. propedeutica teorica della ricerca sociale;
2. propedeutica storica della ricerca sociale;
3. propedeutica metodologica della ricerca sociale;
4. confini epistemologici della sociologia;
5. strumenti teorico-metodologici;
6. tecniche per l'analisi delle processualità micro-sociologiche;
7. tecniche per l'analisi delle processualità macro-sociologiche;
8. prospettive teoriche fondamentali;
9. linguaggio delle scienze sociali;
10. ordine e mutamento;
11. rapporto teoria-ricerca empirica;
12. metodologia e tecnica della ricerca sociale;
13. politica sociale;
14. metodi e tecniche del servizio sociale;
15. sistemi sociali comparati;
16. analisi dei gruppi;
17. analisi della salute;
18. analisi della scienza;
19. analisi dello sviluppo;
20. analisi della sicurezza sociale;
21. metodi della pianificazione;
22. storia del pensiero sociologico.

    …..

Sociologia dei Fenomeni Politici (SPS/11)

1. rapporto fra la società e il mondo delle decisioni strategiche vincolanti;

2. rapporto sistemi sociali-politiche pubbliche;
3. analisi socio-politica;
4. sociologia dell'amministrazione;
5. sociologia delle relazioni internazionali;
6. comunicazione politica

…..

Le categorie individuabili a partire dalle declaratorie dei settori scientifico-disciplinari qui considerati come esempio non presentano tutte lo stesso livello di generalità, né risultano mutuamente esclusive. Qui non si è tentato di affinare o sistematizzare la classificazione, l'esempio è mirato esclusivamente alla presentazione di uno dei possibili esiti, perciò queste caratteristiche sono estremamente evidenti. Inoltre la classificazione non potrebbe essere considerata come esaustiva data l'ineliminabile possibilità dell'emersione di nuove problematiche di studio. Gli elenchi teminano con dei punti di sospensione proprio per sottolineare il carattere esemplificativo e non esaustivo delle classi individuabili nelle declaratorie.

Non rientra tra gli scopi di questo lavoro la proposta di una classificazione tematica per le Scienze Sociali e Politiche, ma si intende evidenziare la possibilità dell'elaborazione di un simile strumento da parte di un gruppo di esperti selezionati *ad hoc* oppure da parte dello stesso GEV. Le declaratorie riferite agli SSD potrebbero rappresentare un buon punto di partenza dato che sono già disponibili per tutte le Aree, inoltre la classificazione contempla già l'affinità tra i settori (Allegato D al DM 4 ottobre 2000), e un simile schema di corrispondenze potrebbe essere messo a punto anche per le meso-categorie. Questa scelta eviterebbe una selezione dei revisori eccessivamente restrittiva, capace di porre a rischio l'opportuna varietà dei punti di vista nella valutazione del prodotto, permettendo allo stesso tempo agli esperti valutatori di poter identificare rapidamente ed efficacemente nell'albo i revisori più adatti alla valutazione di ciascun prodotto. Uno schema simile se non identico potrebbe infatti essere utilizzato anche per la classificazione dei prodotti della ricerca.

Risolta, almeno in parte, la questione dell'individuazione degli ambiti di studio per i revisori resta da sciogliere il problema della qualità dei revisori. La via classica per la selezione dei revisori in ambito editoriale è la reputazione, un criterio utilizzato anche nella VQR, almeno dal GEV14, che tuttavia non risponde ai requisiti di pubblicità e controllabilità che tanto spesso sono stati richiamati.

Una possibilità è includere tutti i candidati a revisore nell'albo per la VQR, facendo leva sulla caratterizzazione *peer* della peer review, l'altra è stabilire dei criteri di selezione, dei requisiti minimi per poter svolgere il ruolo di revisori. Nella stessa scelta dei requisiti è possibile fare appello, nuovamente, alla parità, oppure utilizzare l'eccellenza come criterio di riferimento.

Nell'ottica della parità è possibile accogliere il suggerimento di Krippendorff (1980) per la valutazione dell'affidabilità dei rilevatori nell'analisi del contenuto: sarebbe cioè possibile selezionare come revisori esclusivamente i candidati che non devino eccessivamente dalla media delle valutazioni. Considerare cioè come pari i candidati che formulano giudizi sufficientemente vicini al giudizio medio espresso dall'insieme dei revisori, naturalmente tenendo fermo il riferimento alle categorie tematiche appena discusse. A questo fine sarebbe ovviamente necessario richiedere a tutti i candidati, al momento della candidatura, di valutare almeno un prodotto della ricerca (per ciascuna categoria tematica direttamente connessa alla sua produzione scientifica, o almeno per SSD) utilizzando la stessa scheda di rilevazione progettata per l'esercizio. Una volta rilevati i giudizi dei

revisori sarebbe possibile calcolare i punteggi medi, anche sui singoli criteri, ed escludere i soggetti che si discostino più di una certa soglia (scelta arbitrariamente oppure, ad esempio, pari alla deviazione standard) dalla valutazione media.

Questa prima soluzione ha, ovviamente, dei costi e risentirebbe fortemente di eventuali difetti della scheda di rilevazione (sarebbe, in effetti, essenziale l'utilizzo di una scheda precedentemente validata) o di specifiche caratteristiche dei prodotti utilizzati, tuttavia: «aumentando il numero degli analisti fino a cifre relativamente consistenti […] si può essere più sicuri del fatto che le categorie che verranno scelte saranno, con probabilità crescente in modo proporzionale al numero degli analisti, frutto della condivisione di un determinato codice» (Nobile, 1997, p. 122). In questo caso dunque la media dei punteggi rifletterebbe adeguatamente i criteri di valutazione della comunità scientifica, o meglio di quella parte della comunità scientifica che ha scelto di candidarsi a revisore. Cioè, ad esempio, nell'ipotesi che tutti gli studiosi di alto profilo siano troppo impegnati in altre attività per candidarsi, così come i ricercatori più produttivi, la media dei punteggi rifletterebbe i criteri di valutazione della parte meno prestigiosa della comunità scientifica, rischiando di escludere invece revisori i cui criteri risulterebbero più discriminanti.

Una seconda soluzione dunque potrebbe adottare più o meno la stessa procedura, assumendo però come criterio non la parità, dunque la media, ma l'eccellenza, in questo caso il punteggio assegnato da un revisore ritenuto massimamente competente nel campo. In altri termini si tratta di adottare un'ottica vicina a quella della validità per criterio, utilizzando uno standard ritenuto valido per la valutazione dell'affidabilità dei revisori. Anche questa soluzione si ispira ai controlli dell'affidabilità nell'analisi del contenuto, dove però l'uso di un criterio è riferito allo strumento anziché ai rilevatori[245] (Krippendorff, 1980; Nobile, 1997). Qui la difficoltà principale è la scelta del criterio, cioè del revisore o dei revisori sulla cui base determinare i punteggi da utilizzare come standard. Una possibilità consiste nella selezione da parte del GEV di uno o più revisori esperti per ciascuna categoria tematica, i cui giudizi possano essere assunti come standard; operazione che richiederebbe nuovamente la determinazione di criteri o il riferimento alla reputazione. Un'altra possibilità è sfruttare direttamente il ruolo del GEV, assegnando al panel il compito di valutare i prodotti da utilizzare per la selezione dei revisori e di utilizzare i punteggi risultanti da queste valutazioni collegiali come criteri.

La terza soluzione è l'adozione di standard riferiti direttamente ai revisori, cioè l'individuazione di requisiti minimi cui un candidato deve rispondere per poter svolgere il ruolo di *referee* nel corso della VQR. Mentre le procedure appena esposte si fondano sul controllo *ex-post* dell'affidabilità dei revisori, producendo indicatori espressivi di accordo, l'ultima soluzione fa riferimento a indicatori predittivi di accordo, mirando a individuare *ex-ante* le caratteristiche di un soggetto che ne fanno un buon revisore[246]. I criteri potrebbero essere legati alla produzione scientifica e all'esperienza nell'ambito della valutazione dei candidati, data l'insostenibilità in Area 14, come nelle altre Aree connesse alle scienze sociali, e umane dell'adozione di standard bibliometrici e l'impossibilità di operativizzare altrimenti la reputazione scientifica. Ad esempio,

[245] Si tratta in effetti di una procedura utilizzata di rado, non solo in ragione della difficoltà nel reperire strumenti validati da utilizzare come criterio, ma anche per ragioni tecnico-metodologiche connesse al tipo di variabili da rilevare (Nobile, 1997).

[246] Circa la distinzione tra indicatori espressivi e predittivi si veda Lazarsfeld, 1958 (tr. it. 1967).

stabilendo un criterio per l'esperienza e uno per la produzione scientifica, si potrebbero prevedere i seguenti requisiti per l'iscrizione all'albo dei revisori VQR:

- avere svolto negli ultimi 3 anni almeno una delle seguenti attività:
  - partecipato a comitati editoriali di riviste scientifiche;
  - referaggio per riviste di fascia A;
  - di referaggio per progetti di ricerca (in bandi competitivi nazionali o internazionali).
- per poter essere utilizzato come revisore per la categoria tematica X il candidato deve aver pubblicato un certo numero di prodotti della ricerca[247] nella categoria tematica X.

E' chiaro che una definizione di questo genere non consentirebbe ripensamenti o aggiustamenti nel corso dell'esercizio e che dunque deve essere pensata e perfezionata sulla base delle esigenze dell'esercizio stesso.

In teoria sarebbe possibile a partire dal numero di prodotti da valutare in ciascun settore scientifico-disciplinare, ottenere una stima del numero minimo di revisori necessari e, anche sulla base di questa stima, calibrare i requisiti di selezione dei revisori (ad esempio per determinare la produttività minima per ciascuna categoria tematica). La calibratura dovrebbe in ogni caso avvenire nel rispetto dalla definizione *ex ante* dei criteri. L'obiettivo è soprattutto quello di evitare che la selezione dei revisori avvenga secondo criteri eterogenei ed incostanti nel corso dell'esercizio, come invece è accaduto per la VQR 2004-2010 (*cfr.* ad esempio Anvur, 2013d, GEV14, p.26).

La stima del numero di revisori necessari al compimento della valutazione dei prodotti ha l'ulteriore obiettivo di evitare che i revisori abbiano carichi di lavoro eccessivamente differenti l'uno dall'altro. Si è già sottolineato, infatti, che il numero di prodotti da valutare potrebbe condizionare l'esito delle valutazioni (*cfr*. § 4.1.2), se non per una questione di differenze nelle possibilità di comparazione (Jayasinghe *et al.* 2003) per il tempo a disposizione per ciascuna revisione o per la stanchezza accumulata in relazione alla procedura. Stabilire dei limiti alle valutazioni effettuabili da ciascun revisore potrebbe, se non controllare, quantomeno ridurre il rischio di distorsioni legate a differenze nel carico di lavoro, contribuendo allo stesso tempo a razionalizzare la progettazione *ex ante* dell'esercizio e la selezione dei revisori. Questa scelta potrebbe avere un impatto sulla disponibilità dei revisori a partecipare all'esercizio (ad esempio gli studiosi più impegnati potrebbero scegliere di non partecipare ritenendo eccessivo il carico di lavoro), tuttavia la possibilità di responsabilizzare i revisori chiarendo l'entità dell'impegno richiesto potrebbe limitare il numero di revisioni inevase oppure rifiutate per mancanza di tempo.

Inoltre vale la pena sottolineare l'importanza, in un esercizio di valutazione, dell'uniformità dei profili dei revisori: avere a disposizione una quota di revisori "eccellenti" ridotta e poco disponibile appare più come un fattore di distorsione che come un'assicurazione di qualità della valutazione. I profili dei revisori dovrebbero essere il più uniformi possibili non solo perché i criteri di valutazione di uno studioso inesperto possono essere inadeguati, ma anche perché i criteri di valutazione di un revisore realmente *out standing* potrebbero risultare eccessivamente severi e restrittivi.

[247] Il numero minimo di prodotti potrebbe essere determinato, assumendo un criterio simile a quello utilizzato nell'ASN, utilizzando come soglia il valore mediano del numero di prodotti in ciascuna categoria tematica. Al fine di evitare le complicazioni connesse all'uso di diverse mediane per diversi tipi di prodotti (articoli, volumi, ecc.) sarebbe possibile calcolare un indice di produttività che assegni un peso adeguato a ciascun tipo di prodotto (non semplice da determinare) e utilizzare come soglia il valore mediano della produttività.

Sulla base del numero di soggetti da valutare (professori e ricercatori) è possibile ottenere una stima dei prodotti da valutare per settore disciplinare[248]. A titolo di esempio si riporta una stima del numero di revisori basata sul numero di prodotti attesi e conferiti per SSD in Area 14 nel corso della VQR 2004-2010. La proiezione stima un numero massimo e un numero minimo di revisori immaginando di stabilire il numero minimo di revisioni da effettuare a 10 e il numero massimo a 20 (Tabella 68), si noti che il numero di revisori necessari per la valutazione dei prodotti N/A, cioè non classificabili in SSD, è stato ripartito proporzionalmente tra gli SSD, assegnando però almeno un revisore aggiuntivo a ciascun settore (tutti gli arrotondamenti sono stati effettuati per eccesso).

Il numero minimo di revisori da selezionare (corrispondente a uno scenario in cui tutti i revisori effettuano 20 revisioni) in base ai prodotti attesi sarebbe stato pari a 465, in base ai prodotti conferiti il numero minimo di revisori è invece 448. La selezione iniziale del GEV includeva 443 revisori (Anvur, 2013d, GEV14, p.26), dunque è possibile immaginare che il GEV si attendesse un numero elevato di revisioni per revisore, risulta tuttavia evidente che una più accurata valutazione delle effettive necessità dell'esercizio avrebbe condotto a una selezione iniziale più ampia riducendo almeno in parte le difficoltà emerse nel corso della valutazione dei prodotti.

Tabella 67 - Proiezioni numero minimo e massimo di revisori sul numero di prodotti attesi e conferiti per SSD nella VQR 2004-2010 (dati provenienti dalla Tabella 2.2, Anvur 2013d, GEV14)

SSD	Proiezioni sui prodotti attesi VQR 2004-2010				Proiezioni sui prodotti conferiti VQR 2004-2010			
	Prodotti attesi	Revisioni minime necessarie	Revisori MAX (10 revisioni)	Revisori MIN (20 revisioni)	Prodotti conferiti	Revisioni minime necessarie	Revisori MAX (10 revisioni)	Revisori MIN (20 revisioni)
SPS/01	283	566	60	31	277	554	59	30
SPS/02	361	722	77	39	343	686	72	37
SPS/03	166	332	36	18	155	310	33	17
SPS/04	519	1038	109	55	505	1010	105	53
SPS/05	64	128	14	8	61	122	14	8
SPS/06	165	330	35	18	153	306	33	17
SPS/07	1.046	2092	220	110	1.011	2022	211	106
SPS/08	734	1468	154	78	718	1436	150	75
SPS/09	382	764	81	41	376	752	79	40
SPS/10	184	368	39	20	183	366	39	20
SPS/11	111	222	24	13	108	216	23	12
SPS/12	164	328	35	18	160	320	34	17
SPS/13	83	166	18	10	83	166	18	10
SPS/14	43	86	10	6	42	84	10	6
N/A	189	378	-	-	152	304	-	-
Totale	**4.494**	**8988**	**912**	**465**	**4327**	**8654**	**880**	**448**

In base al numero di revisioni effettuate in ciascun SSD di Area 14 nel corso della VQR (che include sia le revisioni dei prodotti conferiti dai soggetti N/A, sia le terze revisioni, non considerate nella proiezione, riportando le revisioni nell'SSD di valutazione e non per SSD di appartenenza del soggetto valutato), è stato calcolato il numero medio di revisioni da effettuare per revisore (sia in

[248] E' vero che il settore disciplinare di appartenenza dell'autore non determina necessariamente il settore disciplinare in cui il prodotto deve essere valutato ma, sulla base dei dati della VQR 2004-2010, nel secondo esercizio sarebbe possibile elaborare delle proiezioni e dunque apportare aggiustamenti alle stime basate sul solo numero di prodotti attesi.

base al numero massimo che in base al numero minimo di revisori) stimato in base ai prodotti attesi e ai prodotti conferiti (Tabella 68).

Solo in pochi settori disciplinari il numero di revisori stimato non sarebbe risultato sufficiente per il numero di revisioni effettivamente realizzate; si tratta dei casi in cui il numero medio di revisioni effettuate calcolato in base al numero massimo o minimo di revisori stimato, avrebbe superato le soglie previste pari a 10 o 20 revisioni a testa. Questi casi sono esclusivamente riferiti a settori scientifico-disciplinari del Sub-GEV di Scienze Sociali (SPS/8, SPS/9, SPS/11 e SPS/12, evidenziati in grigio in Tabella 68), e lo scostamento dalle soglie risulta rilevante solo per il settore SPS/11 (Sociologia dei fenomeni politici).

Lo scarto tra la media e il massimo di revisioni previste, come ovvio, risulta meno problematico per la proiezione basata sui prodotti attesi. La quota di revisioni stimate per prodotti non conferiti bilancia in effetti, almeno in parte, il numero di terze revisioni effettuate per dirimere i casi *undecided*, corrispondenti ai prodotti con valutazioni divergenti di più di una classe per i primi due revisori.

Tabella 68 – Confronto delle proiezioni del numero minimo e massimo di revisori con le revisioni effettuate per SSD di competenza del revisore nella VQR 2004-2010 (dati provenienti dalla Tabella 2.11, Anvur 2013d, GEV14)

SSD	Revisioni effettuate	Confronto su revisioni effettuate proiezione basata sui prodotti attesi		Confronto su revisioni effettuate proiezione basata sui prodotti conferiti	
		Revisioni per revisore (MAX)	Revisioni per revisore (MIN)	Revisioni per revisore (MAX)	Revisioni per revisore (MIN)
SPS/01	561	9,4	18,1	9,5	18,7
SPS/02	605	7,9	15,5	8,4	16,4
SPS/03	286	7,9	15,9	8,7	16,8
SPS/04	983	9,0	17,9	9,4	18,5
SPS/05	118	8,4	14,8	8,4	14,8
SPS/06	313	8,9	17,4	9,5	18,4
SPS/07	1993	9,1	18,1	9,4	18,8
SPS/08	1578	10,2	20,2	10,5	21,0
SPS/09	819	10,1	20,0	10,4	20,5
SPS/10	389	10,0	19,5	10,0	19,5
SPS/11	277	11,5	21,3	12,0	23,1
SPS/12	348	9,9	19,3	10,2	20,5
SPS/13	145	8,1	14,5	8,1	14,5
SPS/14	97	9,7	16,2	9,7	16,2
Totale	**8512**	**9,3**	**18,3**	**9,7**	**19,0**

Si noti che le simulazioni appena presentate non considerano in alcun modo il tipo di prodotto, che pure è fondamentale ai fini della stima dei carichi di lavoro: la valutazione di tre articoli non richiede lo stesso tempo né lo stesso impegno della valutazione di tre monografie. Come si argomenterà più avanti, questa caratteristica dei prodotti della ricerca andrebbe considerata con la massima attenzione anche e soprattutto al momento della assegnazione dei prodotti ai revisori.

Il perfezionamento della procedura di selezione dei revisori dovrebbe indubbiamente partire dal raggiungimento del massimo numero di candidati possibile, sia a livello nazionale che a livello internazionale. La base volontaria su cui avviene la partecipazione all'esercizio di valutazione può, infatti, rappresentare un fattore di distorsione inevitabile e tutto sommato trascurabile, mentre il mancato coinvolgimento di parte dei soggetti eleggibili a causa di una comunicazione non capillare e/o non tempestiva rappresenta un fattore di distorsione che può e deve essere controllato.

Sarebbe dunque auspicabile che la pubblicazione del bando di partecipazione per i revisori avvenisse con largo anticipo rispetto all'esercizio e che questo bando avesse la massima diffusione possibile coinvolgendo nella comunicazione università, enti di ricerca e società scientifiche a livello nazionale ed internazionale. La finestra temporale per la presentazione dell'auto-candidatura dovrebbe inoltre essere sufficientemente ampia da permettere a tutti i soggetti interessati di ricevere comunicazione del bando e rispondere, ma chiudersi prima dell'apertura dell'esercizio in modo tale da permettere ai GEV di stabilire i requisiti minimi e selezionare i candidati.

Una volta raccolte le candidature e costruito l'albo dei candidati revisori, in base alle esigenze dell'esercizio ma soprattutto in base ai criteri prestabiliti, sarebbe impossibile ridefinire i requisiti minimi nel corso dell'esercizio, come è invece avvenuto nel corso della VQR 2004-2010.

La necessità di coinvolgere *in itinere* revisori non inclusi nell'albo, allo scopo di coprire specifiche competenze, dovrebbe risultare ridotta non solo grazie alla copertura già più ampia dell'albo Anvur, ma anche grazie a questa fase di progettazione *ex ante*. Una maggiore varietà dei profili, a parità di valore scientifico, permetterebbe agli esperti valutatori di gestire più facilmente i conflitti di interesse e di considerare nell'assegnazione dei prodotti anche le caratteristiche rilevanti di revisori e autori[249].

Il protocollo di assegnazione dei prodotti utilizzato nella VQR 2004-2010 era in grado di controllare alcune caratteristiche, ad esempio il ruolo accademico, l'affiliazione istituzionale, ed i legami di co-autoraggio. Il controllo di queste caratteristiche prima dell'assegnazione riduce sia il numero di rifiuti delle revisioni a causa di conflitti di interesse sia il rischio di distorsioni, facilitando allo stesso tempo il compito dei membri del GEV. Il protocollo, purtroppo, non è mai stato reso noto, nonostante la pubblicazione migliorerebbe sensibilmente il livello di pubblicità e replicabilità della procedura, riducendo il margine di critica in relazione almeno a questo aspetto.

Sarebbero inoltre da controllare eventuali differenze nella varietà dei prodotti per i singoli revisori, o nella varietà dei revisori per i prodotti di uno stesso autore. Nel primo caso sarebbe opportuno far sì che a ciascun revisore vengano assegnati prodotti di diversi autori, affiliati a diverse istituzioni e con un diverso ruolo accademico che riguardino tematiche differenti, per quanto attinenti al campo di competenza del revisore stesso. Evidentemente si tratterebbe di una serie complessa di controlli che difficilmente potrebbero essere effettuati *in itinere* senza incidere sui tempi di realizzazione dell'esercizio, sembra però centrale l'adozione di un simile sistema almeno con riferimento al tipo di prodotto. Non è possibile immaginare che un revisore impieghi lo stesso tempo e la stessa attenzione nel revisionare un articolo o una monografia, dunque sempre nella direzione della massima uniformazione possibile del carico di lavoro sarebbe utile implementare dei controlli già nella fase di assegnazione dei prodotti.

Un altro suggerimento riguarda la varietà dei revisori per i prodotti di uno stesso autore: allo scopo di ridurre l'effetto di eventuali distorsioni nel giudizio del revisore nei confronti del soggetto o del suo ente di appartenenza a ciascun revisore dovrebbe essere assegnato un solo prodotto per autore. Si tratta naturalmente di piccoli accorgimenti, tuttavia ciascuno di essi elimina o riduce la possibilità che la valutazione dei prodotti risulti inaffidabile.

[249] Una questione evidenziata dagli EV nelle interviste: «il problema è che, ripeto, c'erano pochi valutatori, pochissimi su alcuni argomenti, e molto spesso in conflitto. Perché quando sono pochissimi, sono quelli… in alcune discipline in cui i valutatori erano pochissimi perché non c'è n'erano altri. Questo ha creato conflitti di interessi, oppure conflittualità interna» (Intervista Bazzicalupo).

Infine, circa l'assegnazione dei prodotti, sarebbe di grande interesse la progettazione e la messa a punto delle procedure in grado di facilitare il compito degli EV classificando i contenuti dei prodotti secondo lo stesso schema usato per l'individuazione delle aree di competenza dei revisori, stabilendo eventualmente delle regole di massima per assegnazione. L'adozione di una classificazione a più livelli che faccia riferimento alle declaratorie degli SSD potrebbe permettere agli autori di segnalare direttamente alcune categorie tematiche per ciascun prodotto, segnalando anche in quale categoria il prodotto andrebbe valutato. In questo modo prodotti e revisori condividerebbero una stessa classificazione tematica, derivante da quella in uso (gli SSD), ma più particolareggiata, con evidenti vantaggi pratici con riferimento alla fase di assegnazione dei prodotti.

Un sistema di questo genere non dovrebbe essere eccessivamente rigido, se così fosse infatti si stabilirebbero dei confini netti e stabili che difficilmente troverebbero riscontro nella pratica scientifica. In questa fase il ruolo degli EV è essenziale, tuttavia un sistema di classificazione renderebbe meno complessa e più trasparente la fase di assegnazione dei prodotti ai revisori.

Una possibilità ulteriore è la creazione di un sistema di parole chiave che siano legate sia ai contenuti prodotti che all'esperienza scientifica dei revisori e riconducibili a categorie di ampiezza via via maggiore. Sulla base di questo sistema di classificazione sarebbe possibile identificare con minore difficoltà i revisori con le competenze più adatte alla valutazione di ciascun prodotto, stabilendo *ex ante* delle regole di corrispondenza o anche solo fornendo queste informazioni agli esperti valutatori che hanno il compito di assegnare il prodotto.

La creazione di un simile strumento tuttavia non è affatto semplice e richiederebbe una enorme mole di lavoro. Si pensi, a titolo di esempio, a una delle classificazioni tematiche più note nel campo delle scienze sociali, quella in uso per i contenuti degli articoli nel database *Sociological Abstracts* (di ProQuest). L'indicizzazione degli articoli in *Sociological Abstracts* lavora su tre livelli:

1. i *Classification Codes*, che indicano il principale soggetto dell'articolo, contemplano 19 aree, abbastanza ampie, e 95 sottovoci più specifiche[250];
2. i *Descriptors*, assegnati tramite il *Thesaurus of Sociological Indexing Terms* (attualmente alla sesta edizione);
3. gli *Identifiers,* vocaboli che riflettono concetti nuovi o emergenti non ancora aggiunti non al Thesaurus.

Questo sistema, naturalmente, è in continua evoluzione e presenta innumerevoli problematiche connesse all'esaustività e alla mutua esclusività delle categorie (si veda ad esempio Pathak, 2000). Basta osservare l'elenco delle sottovoci per individuare uno sbilanciamento della classificazione a favore dei temi piuttosto che dei metodi o degli approcci, che pur non essendo del tutto assenti risultano solo parzialmente rappresentati: ad esempio nessuna delle sotto-voci dell'area connessa alla metodologia fa esplicitamente riferimento a tecniche non standard, e mentre un'intera area è dedicata alla sociologia critica non vi è traccia di riferimenti ad altri approcci come l'interazionismo o il funzionalismo.

E' vero che una parziale semplificazione potrebbe scaturire dalla creazione di tre classificazioni distinte: una per le tematiche, una per gli approcci teorici, una per le tecniche, ma non avvicinerebbe affatto a una soluzione. In un contributo sulla classificazione delle tecniche di ricerca nelle scienze sociali Durrant (2009) ha evidenziato come sia possibile scegliere diversi *fundamenta*

[250] La classificazione è disponibile on-line: http://proquest.libguides.com/ld.php?content_id=3028061.

divisionis e selezionare in base a diversi criteri gli elementi da includere o escludere dalla classificazione, rendendo evidente come le categorie risultino difficilmente mutuamente esclusive.

Nel caso della VQR la soluzione più conveniente potrebbe essere rappresentata da una classificazione che parta dall'esistente, cioè dalle parole chiave segnalate dagli autori al momento della sottomissione dei prodotti e dagli SSD[251]. Adottando dunque un approccio *grounded*, le parole chiave segnalate dagli autori potrebbero essere ricondotte allo stesso schema classificatorio utilizzato per i campi di *expertese* dei revisori, senza necessariamente rispettare i requisiti classici della classificazione, permettendo dunque che una stessa parola chiave sia collegata a più voci o anche a nessuna di esse nel caso in cui nuove tematiche o nuovi approcci non risultino direttamente riconducibili allo schema elaborato. E' tuttavia evidente, date le complicazioni di questa soluzione, il vantaggio presentato da una procedura che preveda delle categorie tematiche da parte degli autori.

Una volta razionalizzata la procedura di selezione dei revisori e di assegnazione dei prodotti l'attenzione va posta sugli stessi revisori. E' stato già sottolineato quanto la competenza nel campo non sia l'unico prerequisito di una valutazione affidabile: spesso è necessario che i revisori acquisiscano competenze specifiche a seconda degli obiettivi della valutazione. Nel caso della VQR le informazioni preliminari fornite ai revisori circa la valutazione erano prettamente tecniche e riguardavano la procedura informatica di compilazione e invio della scheda, ma a questo genere di informazioni andrebbe affiancata una formazione specifica mirata innanzitutto alla presentazione dei criteri di valutazione (si pensi agli evidenti problemi di fraintendimento del criterio relativo all'internazionalizzazione), poi alla presentazione della scheda di valutazione e alle procedure di assegnazione della classe di merito, e infine alla responsabilizzazione dei revisori[252].

Una chiara esposizione dei criteri di valutazione e degli obiettivi dell'esercizio potrebbe almeno in parte ovviare ai problemi evidenziati in relazione alla vaghezza e all'ambiguità delle definizioni ministeriali, dando modo ai revisori di riflettere su ciascun criterio e sulle caratteristiche che un prodotto deve presentare per soddisfarlo, moderando le differenze nell'interpretazione di criteri e classi di merito.

La presentazione della scheda di rilevazione, dei suoi contenuti semantici e degli esiti della sua compilazione renderebbe i revisori più consapevoli dell'esito della propria valutazione, limitando nuovamente i fraintendimenti circa l'assegnazione dei punteggi e delle classi di merito e sensibilizzando circa la necessità di una adeguata compilazione del campo relativo alle motivazioni. Tanto più i revisori saranno informati circa i presupposti e gli esiti della compilazione delle schede, tanto meno i dati rilevati dovrebbero rischiare di subire distorsioni in questa fase dovute a distrazioni o fraintendimenti nella compilazione. Inoltre questa formazione preliminare priverebbe del suo

[251] In Italia infatti l'ultimo *thesaurus* di sociologia risale alla fine degli anni '90 (1999), e senza dubbio necessiterebbe di un aggiornamento e un ampliamento per poter essere utile allo scopo, inoltre non sono mai state elaborate classificazioni tematiche specifiche.

[252] Questioni citate in alcune interviste: «un miglioramento che potrebbe essere introdotto è la possibilità di formare i *referee*, cioè di introdurre linee guida e di fare dei tutorial, anche web-based, perché si chiariscano gli elementi di professionalità richiesti per svolgere questo ruolo» (Intervista Bonaccorsi); «intendiamo costruire un manuale per i revisori più accurato» (Intervista Benedetto); «la responsabilità dei valutatori dovrebbe essere un po' più chiara, perché questa non c'è. Purtroppo non ci sta per via del fatto che il meccanismo è occulto e tutti sono… però un sistema di responsabilizzazione ci dovrebbe essere, magari da parte di colui che fa l'attribuzione e risulterà responsabile dei giudizi» (Intervista Bazzicalupo).

scopo il controllo *ex post* da parte del revisore della classe di merito assegnata, annullando i rischi di distorsione che ne conseguivano.

Infine sarebbe opportuno esplicitare ai revisori le loro responsabilità, stilando una sorta di codice deontologico che ciascun revisore sarebbe tenuto a sottoscrivere per poter partecipare all'esercizio[253]. Ad esempio ciascun revisore dovrebbe essere tenuto a valutare il prodotto in base al suo contenuto (senza dunque considerare le informazioni, pure disponibili su autori e collocazioni editoriali), considerandolo nello specifico contesto della disciplina di riferimento, e dunque valutando non le tecniche o gli approcci di riferimento, ma il loro utilizzo in relazione agli obiettivi cognitivi, la loro adeguatezza agli scopi, la coerenza dell'argomentazione. La valutazione dovrebbe essere basata sugli stessi requisiti della comunicazione scientifica: onestà, integrità, accuratezza. Ciascun prodotto dovrebbe essere valutato con la stessa attenzione, indipendentemente dal suo volume o dal livello d'interesse del revisore. Un codice di questo genere non farebbe che elencare le regole su cui si basano tutte le procedure di peer review, ma avrebbe il vantaggio di renderle esplicite e ben presenti nella mente del revisore al momento della valutazione.

Questa formazione, per quanto breve, per evidenti motivi economici e organizzativi non potrebbe avvenire coinvolgendo di persona tutti i revisori, nondimeno sarebbe pensabile la strutturazione di uno o più moduli di auto-formazione on-line, accompagnati da documenti sintetici e chiari, che sarebbe possibile pubblicare sulla stessa piattaforma utilizzata per la valutazione e che potrebbero essere resi propedeutici alla partecipazione all'esercizio.

6.2.1.1 L'approfondimento delle questioni legate all'uniformità e alla stabilità delle scale di giudizio dei revisori

La centralità delle problematiche che si riferiscono all'uniformità e alla stabilità delle scale di giudizio dei revisori rende di estremo interesse un approfondimento mirato. La letteratura, infatti, presenta diversi studi circa il grado di concordanza dei giudizi dei singoli revisori (Scott, 1974; Scarr e Weber, 1978; Crandall, 1978; Cole *et al.* 1981; Cicchetti, 1991; Oxman *et al.* 1991; Callaham *et al.* 1998), tuttavia in genere le analisi sono riferite alla peer review propedeutica alla pubblicazione e i risultati presentati risultano discordanti e solo raramente incoraggianti (per tutti Lindsey, 1988 e Hojat *et al.* 2003).

Considerando anche il dibattito circa l'auspicabilità di una non completa concordanza, legata al rischio di ridondanza dei giudizi (Cole *et al.* 1981; Balair, 1991; Bornmann, 2008), e quello del legame tra la mancanza di accordo e l'appartenenza a differenti posizioni teoriche o metodologiche (paradigmi) o a differenze nelle aree di competenza tra i revisori (Chubin e Hackett, 1990; Eckberg, 1991; Kostoff, 1995; Laudel; 2006) la questione diventa di estremo interesse non solo in relazione alla valutazione della ricerca, ma anche con riferimento alla sociologia della scienza.

Il rischio di ridondanza e l'influenza delle posizioni paradigmatiche dei revisori sui giudizi rivestono un ruolo centrale con riferimento alle scienze sociali proprio in relazione alle loro peculiarità (in termini di relazioni tra specializzazioni, approcci metodologici e teorici, d'interdisciplinarietà, di collaborazione scientifica e di consenso; Moody, 2004), dunque

[253] A questo proposito sono disponibili diversi contributi, per lo più riferiti alla peer review per le riviste; si veda ad esempio Rockwell, 2005.

richiederebbero un approfondimento, tramite l'analisi dei dati disponibili (potrebbero essere utilizzati i dati della VQR 2004-2010) oppure la realizzazione di studi *ad hoc*, magari d'impronta (quasi)sperimentale.

L'analisi dei dati della VQR 2004-2010 potrebbe offrire diversi spunti interessanti, costituendo una base evidenziale di notevole ampiezza, pur non essendo stata costruita *ad hoc*. Sarebbe ad esempio possibile esaminare le quote di giudizi concordanti in relazione al tipo di prodotto, alla lingua di pubblicazione, alle caratteristiche degli autori (genere, settore disciplinare, ruolo accademico e afferenza istituzionale, solo per citare quelle di maggiore interesse). Allo stesso modo sarebbe possibile analizzare il grado di accordo in base all'omogeneità/eterogeneità dei revisori su una serie di caratteristiche: ad esempio l'accordo è maggiore quando i due revisori hanno lo stesso ruolo accademico o appartengono allo stesso settore disciplinare?

Uno studio di questo genere permetterebbe di individuare eventuali anomalie, ad esempio settori disciplinari in cui l'accordo è estremamente basso o estremamente alto, oppure caratteristiche dei prodotti o dei revisori apparentemente in grado di influire sul grado di accordo. Naturalmente non s'intende assumere il grado di accordo come un indicatore di qualità della valutazione, lo scopo qui è individuare e studiare i casi in cui il livello di accordo risulta significativamente differente da quello atteso e di cercare di comprendere le ragioni di questa deviazione dalla norma. La logica sottostante è quella dell'analisi dei casi devianti di lazarsfeldiana memoria: approfondire i casi che si discostano dalla regola per poter specificare la regola stessa (Boudon e Lazarsfeld, 1966, tr. it. 1969).

L'analisi sarebbe anche di maggiore interesse nel caso fossero disponibili i dati sui singoli criteri oltre che le classi di merito assegnate. Si è già rilevato (*cfr*. § 4.3.2) che il grado di accordo tra i revisori al livello dei singoli criteri non è stato considerato nelle procedure di assegnazione delle classi di merito, né sono disponibili dati o evidenze su questa specifica questione. Una base empirica che includa i punteggi assegnati ai prodotti sui singoli criteri permetterebbe di approfondire significativamente lo studio delle scale di giudizio dei revisori, permettendo di analizzare il grado di accordo dei revisori criterio per criterio, e di individuare eventuali differenze nell'uso delle scale per i giudizi analitici. Sarebbe inoltre possibile analizzare il grado di accordo tra il giudizio espresso sui singoli criteri e la classe di merito finale.

Un pre test delle schede di valutazione o un pre test mirato a rilevare il grado di accordo tra i revisori (*cfr*. § 6.1.1) offrirebbero senz'altro l'occasione di studiare più approfonditamente la questione dell'uniformità delle scale di giudizio. I dati prodotti in questa fase, infatti, sarebbero riferiti a *n* prodotti identici per tutti i revisori, o nel caso di un pre test della scheda per revisori selezionati sulla base di caratteristiche rilevanti (ruolo accademico, anzianità di ruolo, settore disciplinare, afferenza istituzionale). Un pre test condotto con una procedura standard sarebbe ampiamente sufficiente per l'analisi del grado di accordo tra i revisori, tuttavia sfruttando anche l'*in depth probes procedure* vi sarebbe modo di comprendere come i *referee* interpretano i criteri e utilizzano i punteggi, di identificare più chiaramente i casi in cui quest'uso risulta difforme e, eventualmente, di apportare dei correttivi mirati al fine di massimizzare l'uniformità delle scale di giudizio. In entrambi i casi se i prodotti sono stati selezionati opportunamente (se cioè presentano caratteristiche diverse e risultano variamente classificabili per la loro qualità) analizzando e confrontando le distribuzioni dei punteggi sarebbe possibile identificare la tendenza a utilizzare solo

una parte della scala, ad assegnare punteggi estremi o mediani, ecc. per i singoli revisori, e studiare le eventuali relazioni tra queste tendenze e le caratteristiche dei revisori o dei prodotti.

Né l'analisi dei dati della VQR né quella di un eventuale pre test permetterebbero tuttavia di controllare la stabilità delle scale di giudizio dei revisori. Solo la conduzione di uno o più studi *ad hoc* potrebbe, infatti, produrre una sufficiente base evidenziale. Studi di questo genere potrebbero essere estremamente fruttuosi anche in relazione all'uniformità di giudizio, soprattutto se condotti in un'ottica (quasi) sperimentale.

Si pensi innanzitutto alle differenze tra una revisione dei pari che avviene dopo la pubblicazione e la classica peer review cui i prodotti della ricerca vengono sottoposti prima di essere pubblicati: l'influenza della collocazione editoriale sul giudizio, della sostanziale impossibilità dell'oscuramento degli autori, dell'eventualità che il revisore conosca già il lavoro da valutare, sono solo alcuni dei punti che meriterebbero un approfondimento.

Il fatto che sempre più riviste italiane stiano adottando sistemi di referaggio strutturati e tracciabili, anche per effetto della crescente rilevanza di questo fattore per le procedure di valutazione come la VQR e l'ASN, costituisce una condizione particolarmente favorevole allo sviluppo di questo campo di studi.

Nelle scienze politiche e sociali in particolare, ma anche in tutti gli altri ambiti scientifici, si avrebbe l'occasione di mettere a fuoco le conseguenze sulla valutazione di differenze più o meno marcate tra gli approcci teorici e metodologici o i campi di studi di autori e revisori. Inoltre sarebbe estremamente interessante poter studiare la relazione tra la qualità della revisione (in termini di accuratezza, correttezza e onestà) e caratteristiche come l'esperienza di ricerca e di revisione dei *referee*.

Lo studio della stabilità e dell'uniformità delle scale di giudizio non è l'unica ragione per cui la realizzazione di studi mirati e ben strutturati sarebbe auspicabile. L'interesse va al di là del controllo dei possibili fattori di distorsione nel corso della VQR e investe tutta una serie di tematiche di grande interesse per la sociologia della scienza; le basi e gli standard in base ai quali i pari valutano un prodotto scientifico sono connessi non solo con la valutazione istituzionale o l'accesso alla pubblicazione, ma anche con la costruzione e il mantenimento del consenso, l'*ethos* della scienza, la comunicazione e la condivisione dei risultati della ricerca.

Varrebbe davvero la pena di riconnettere il discorso sulla peer review a quadri teorici e dibattiti più ampi, poiché, come evidenziato da Gläser e Laudel (2006), dopo Merton la letteratura sulla peer review è diventata sempre più a-teoretica ed empiristica, centrata essenzialmente sulla validità e l'attendibilità delle procedure. Un contributo interessante in questa direzione è stato apportato da Bornmann (2011), con una rassegna della ricerca sulla peer review mirata a connettere le evidenze empiriche disponibili ai principali approcci alla sociologia della scienza: quello della scuola nord-americana, quello costruttivista e quello connesso alla teoria dei sistemi sociali di Luhmann.

6.2.2 Proposte per la valutazione bibliometrica

Le uniche fonti di dati utilizzabili per l'analisi bibliometrica della produzione scientifica sono i database citazionali commerciali il cui scopo originale, come si è già sottolineato più volte, non era la valutazione della ricerca, ma la valutazione delle riviste per scopi biblioteconomici.

La creazione e la gestione di archivi citazionali *ad hoc* non sembrano percorribili dati gli elevatissimi costi in termini di denaro, lavoro e tempo, di una simile operazione. Sorgerebbero inoltre innumerevoli questioni circa i criteri di selezione delle riviste da indicizzare, la copertura del nuovo database dal punto di vista geografico e temporale, gli algoritmi di tracciamento e conteggio delle citazioni e la classificazione dei documenti e delle riviste.

Volendo utilizzare l'analisi bibliometrica per la valutazione dei prodotti della ricerca è dunque necessario affidarsi ai database esistenti, ma ciò non vuol dire necessariamente subirne tutti i limiti. Ad esempio il CWTS (*Centre for Science and Technology Studies*) di Leiden, dagli inizi degli anni '90, utilizza per i suoi studi una versione "corretta" del database ISI (WoS), al fine di ottenere un dato il più affidabile possibile circa le affiliazioni degli autori o le loro nazionalità (Moed *et al.* 1995; Waltman *et al.* 2012). Il database CWTS nelle sue ultime versioni ha visto l'introduzione di una serie di innovazioni connesse agli obiettivi di ricerca dell'Istituto (ad esempio esclude dal calcolo le auto-citazioni, perché assumono un peso variabile nel determinare i valori degli indicatori a seconda delle istituzioni), ciò nonostante gli studi del gruppo di Leiden, incluso il *Leiden Ranking*, risentono ancora dei limiti connessi alle caratteristiche dei database (Waltman *et al.* 2012).

Non è dunque impossibile la messa a punto di versioni "corrette" dei database citazionali sulla base degli obiettivi della VQR, cercando ad esempio di effettuare dei controlli sulla classificazione dei documenti in tipi oppure mettendo a punto e utilizzando una classificazione tematica *ad hoc*, ma anche queste azioni avrebbero un costo notevolissimo a fronte di un esito incerto.

Qualsiasi azione mirata al miglioramento dell'affidabilità e della validità delle basi di dati disponibili dovrebbe essere preceduta da uno studio approfondito, e comparativo, non solo della struttura dei database, degli algoritmi di individuazione e conteggio delle citazioni, degli identificativi di riviste, autori, istituzioni, ma anche della questione concernente la copertura della letteratura e alla rispondenza delle riviste indicizzate alle caratteristiche della *core literature* e infine delle conseguenze di questo insieme di connotati sulla qualità dei dati citazionali.

Dato il genere di uso che si è fatto nel corso della VQR dell'analisi bibliometrica, come nel caso della peer review è opportuno approfondire diverse problematiche relative all'accordo tra i revisori, nel caso della bibliometria sarebbe auspicabile l'approfondimento delle questioni relative alla sovrapponibilità e alle differenze tra i principali database citazionali (*cfr*. § 4.2). Un approfondimento che richiederà del tempo, dato che la stessa comunità scientometrica sta ancora elaborando nuove strategie in base ai cambiamenti avvenuti nel panorama della disponibilità e della gestione dei dati citazionali (Glänzel e Moed, 2008).

Dal punto di vista della valutazione della ricerca, tuttavia, anche piccoli studi sulle conseguenze che le più o meno rilevanti differenze tra i database possono avere sulla valutazione dei prodotti potrebbero dare i loro frutti, fornendo almeno gli strumenti per definire procedure più attendibili.

Una proposta sicuramente attuabile, che risulta anche economica sotto diversi punti di vista, è relativa all'utilizzo di un solo database bibliometrico per disciplina, non necessariamente di un solo database per l'intero esercizio. Molte delle possibili distorsioni individuate sono legate proprio all'uso di più database citazionali e alle loro differenze in termini di copertura, struttura e funzionamento, e potrebbero essere evitati semplicemente scegliendo un solo database di riferimento ad esempio per Area, Sub-GEV o SSD. Il problema fondamentale è che la valutazione ottenuta può variare a seconda del database, dunque, anche scegliendo un criterio univoco, i prodotti indicizzati in uno solo dei database potrebbero essere penalizzati o avvantaggiati proprio grazie alla loro assenza nell'altro.

L'entità del vantaggio o dello svantaggio, ancora una volta, dipenderebbe dal campo di studi, dal genere di documento, dal genere di rivista, e non è valutabile allo stato attuale della conoscenza sulla sovrapponibilità e le differenze tra WoS e Scopus.

L'uso di più database, ciascuno connesso a specifici campi disciplinari, garantirebbe comunque l'autonomia dell'Agenzia rispetto ai fornitori di dati, senza però aumentare i rischi di distorsione. Già nel corso della VQR l'Area 5 (Biologia) e l'Area 6 (Medicina) hanno utilizzato esclusivamente WoS, dunque non è da escludere la possibilità di affidare ai GEV la scelta del database di riferimento. Quest'opzione non creerebbe problemi di comparabilità tra Aree, poiché: «tra le finalità della VQR non compare il confronto della qualità della ricerca tra aree scientifiche diverse» (Anvur, 2013a, p. 7). Inoltre anche diversi Sub-GEV hanno utilizzato criteri e procedure diverse all'interno di alcune Aree, a volte adottando procedure differenti per SSD (ad esempio l'Area 1, Matematica e scienze informatiche), per questa ragione l'Anvur segnala che: «mentre in alcuni casi è possibile confrontare la qualità della ricerca tra SSD della stessa Area, in altri casi (evidenziati nei singoli rapporti di Area) tale confronto non è possibile né opportuno» (*ibidem*).

A livello metodologico sarebbe più corretto scegliere un database per ciascuna disciplina, sulla base del livello di copertura dei prodotti da valutare o del grado di fiducia della comunità scientifica, dunque gli SSD apparirebbero come la scelta migliore. Ciò si tradurrebbe però in una comparabilità davvero ristretta delle valutazioni.

Decidere invece di utilizzare un solo database per ciascun'Area sarebbe perfettamente argomentabile per alcune di esse (ad esempio Medicina o Biologia, cui afferiscono discipline diverse ma omogenee tra loro), molto meno sostenibile e adeguata all'obiettivo della comparazione tra elementi simili nel caso di Aree come Matematica e Scienze dell'Informazione, Ingegneria Civile e Architettura, Scienze Economiche e Statistiche, caratterizzate da un'elevata eterogeneità interna e dall'inclusione di disparate comunità scientifiche.

L'opzione più corretta appare dunque la scelta di un solo database di riferimento per ciascun Sub-GEV, cercando di individuare un punto di equilibrio tra la necessità di utilizzare procedure comparabili e la necessità di ottenere dati affidabili.

Circa l'affidabilità dei dati e l'adeguatezza degli strumenti offerti dai database è importante ribadire la centralità del ruolo di stakeholders che agenzie di valutazione nazionali come l'Anvur possono assumere. Questo ruolo potrebbe condurre a un nuovo sviluppo dei database citazionali e invertire le attuali tendenze a un ampliamento della copertura della letteratura non in grado di assicurare la centralità dei documenti indicizzati (Glänzel e Moed, 2008).

Un ultimo suggerimento che va avanzato riguarda l'effettiva implementazione di una *informed* peer review piuttosto che di una valutazione diretta tramite analisi bibliometrica. Non s'intende qui proporre l'adozione di una procedura di revisione dei pari che ricalchi il sistema adottato per i prodotti privi di indicatori bibliometrici, ma una procedura che assegni ai gruppi di consenso interni ai GEV maggiori responsabilità.

Tra i passaggi procedurali che richiederebbero il controllo attivo di valutatori esperti nel campo, va senza dubbio annoverata l'assegnazione dei prodotti alle *subject categories*, sia nel caso questa sia stata segnalata dall'autore o dalla struttura, sia nel caso vada assegnata *ex novo*.

La centralità del GEV, tuttavia, come nel caso della procedura di valutazione dei prodotti in peer review sarebbe da riferire all'esito della valutazione di ciascun prodotto. Una discussione attiva di tutte le classificazioni all'interno dei gruppi di consenso, al fine di validare la classe di merito

ottenuta dal prodotto, costituirebbe un passaggio di controllo degli esiti della valutazione bibliometrica aumentando la probabilità di individuare e correggere, per quanto possibile, errori o distorsioni di qualsiasi genere. Infatti: «l'uso dell'analisi citazionale dovrebbe basarsi sull'idea che l'impatto citazionale, nonostante gli aspetti più utili e preziosi, non coincide pienamente con nozioni come l'influenza intellettuale, il contributo al progresso scientifico o la qualità della ricerca[254]» (Moed, 2008, p. 161) ed è per questa ragione che «anche le più avanzate tecniche bibliometriche hanno bisogno di essere spiegate in un contesto di revisione tra pari. Così queste tecniche supportano anziché minacciare il ruolo dei pari[255]» (van Raan, 1996, p. 421).

La prima regola nell'utilizzo della bibliometria nella valutazione della ricerca è che «la valutazione quantitativa dovrebbe supportare la valutazione qualitativa da parte di esperti. Le metriche quantitative possono contrastare le tendenze discorsive nella peer review e facilitare la deliberazione. Ciò dovrebbe rafforzare la valutazione tra pari, perché dare giudizi sui colleghi è difficile senza una serie di informazioni rilevanti. Tuttavia, valutatori non devono essere tentati di cedere il processo decisionale ai numeri. Gli indicatori non devono sostituirsi a giudizio informato. Ognuno mantiene la responsabilità delle proprie valutazioni[256]» (Hicks *et al.* p. 420).

Una modifica di questo genere alle procedure in uso comporterebbe dei costi in termini di impegno e di tempo ma, considerando anche gli effetti che la riduzione della distanza tra la finestra temporale di riferimento e la realizzazione dell'esercizio avrà sull'affidabilità dei dati, una più ampia discussione nell'ambito dei gruppi di consenso appare, più che opportuna, necessaria.

[254] Traduzione dall'originale in lingua inglese.
[255] Traduzione dall'originale in lingua inglese.
[256] Traduzione dall'originale in lingua inglese.

Capitolo 7
Oltre la procedura: gli obiettivi della valutazione

7.1 Lo scopo della VQR

La Valutazione della Qualità della Ricerca risponde a uno dei compiti fondamentali assegnati all'Anvur nell'atto della sua istituzione: valutare «la qualità dei processi, i risultati e i prodotti delle attività di gestione, formazione, ricerca, ivi compreso il trasferimento tecnologico delle università e degli enti di ricerca» (DPR 76 del 1/2/2010; art. 3, comma 1, lettera a).

Nella premessa al rapporto finale vengono esposte chiaramente e schematicamente le finalità della VQR 2004-2010:

- «presentare al Paese una valutazione imparziale e rigorosa della ricerca nelle università, negli enti di ricerca e nelle loro articolazioni interne (dipartimenti, istituti,…), che ognuno potrà utilizzare per i propri scopi:
    - gli organi di governo delle Strutture per intraprendere azioni volte a migliorare la qualità della ricerca nelle aree che appaiono deboli rispetto al panorama nazionale;
    - le famiglie e gli studenti per orientarsi nelle difficili scelte collegate ai corsi di studio e alle università[257];
    - i giovani ricercatori per approfondire la propria formazione e svolgere attività di ricerca nei migliori dipartimenti;
    - le industrie e gli enti pubblici per indirizzare la domanda di collaborazione alle strutture che ospitano, nelle aree scientifiche di loro interesse, gruppi di ricerca validi per qualità e massa critica»;
- «determinare una graduatoria nazionale per area scientifica e per struttura basata sugli indicatori del Bando che costituisca uno degli elementi su cui basare la distribuzione della quota premiale del Fondo di Finanziamento Ordinario delle università»;
- «offrire una valutazione dei dipartimenti degli atenei e delle sottostrutture degli enti di ricerca agli organi di governo interni per orientare, nella loro autonomia, la distribuzione interna delle risorse acquisite»;

[257] Poco oltre si precisa che: «la valutazione e i risultati qui descritti non riguardano in alcun modo la qualità e quantità dell'attività didattica che si svolge nelle università. L'Anvur ritiene comunque che una buona didattica richieda, a ogni livello, la presenza di un'attività di ricerca adeguata. Quindi il rapporto può essere utile anche per orientare le scelte dei giovani, particolarmente laddove la ricerca gioca un ruolo importante, vale a dire per i corsi di laurea magistrale e soprattutto per i corsi di dottorato» (Anvur, 2013a, p. 7).

- «consentire un confronto della qualità della ricerca nazionale con quella dei principali paesi industrializzati» (Anvur, 2013a, p. 6).

In sintesi sono individuabili almeno quattro obiettivi distinti: (a) l'informazione degli *stakeholders*; (b) la distribuzione delle risorse a livello nazionale; (c) la distribuzione delle risorse a livello delle strutture; (d) il confronto con la qualità della ricerca internazionale.

Le parole del Presidente Anvur, Stefano Fantoni, individuano questi stessi obiettivi in relazione all'uso dei dati prodotti dalla VQR: «Uno: il Ministero, avuti questi dati, deve usarli per la premialità, li usi come vuole perché è lui che deve farlo. Due: un uso nella *governance* nell'Università […] aiutare un processo di autovalutazione vero dell'università. Tre: un livello informativo, nel senso che se un ragazzo vuole iscriversi a fisica va a vedere dov'è la struttura migliore in Italia […] Questi erano i tre livelli informativi, le tre missioni che noi volevamo adempiere e un quarto obiettivo, più generale e più globale, era mettere l'Italia in corsa, al pari delle altre università europee» (Intervista Fantoni).

Il primo obiettivo, dal punto di vista dell'Agenzia, è stato ampiamente raggiunto: «con l'effetto ranking, che ha generato un effetto mediatico straordinario, anche in parte inatteso […] Da questo punto di vista l'informazione ha circolato molto, meno la parte analitica perché l'abbondanza di informazioni richiede poi del tempo per essere processata, però diciamo questo è stato ottenuto» (Intervista Bonaccorsi). Non senza criticità: «dai media è venuto fuori, forse anche per colpa comunicativa nostra, non lo so, una forzatura a tutto questo perché hanno voluto fare una classifica […] un uso secondo me un po'troppo non opportuno da parte dei media, che naturalmente è difficile tenere a bada» (Intervista Fantoni).

Il secondo, cioè l'obiettivo di costituire una base informativa per l'allocazione delle risorse, è in effetti una realtà: «il secondo effetto si sta ottenendo, nel senso che la prima quota premiale si sta già ottenendo, è stata allocata, adesso siamo alla seconda» (Intervista Bonaccorsi). Nel 2013 il 13,5% delle risorse viene distribuito per la premialità, la quota decisa dai risultati della VQR corrisponde al 90% del 66% distribuito in base alla qualità della ricerca, per un totale di 486.486.000 euro (DM 20 dicembre 2013, n. 1051). L'importanza dei risultati della VQR aumenta sensibilmente nel 2014, sia in rapporto al totale della quota premiale sia in ragione della crescita relativa della quota di fondi distribuiti in base alla premialità. Nel 2014, infatti, la quota premiale degli FFO corrisponde a circa il 18% delle risorse disponibili, ed è decisa per il 70% (pari a 850.500.000 euro) sulla base dei risultati ottenuti nella VQR (DM 4 novembre 2014, n. 815). Nelle parole del coordinatore nazionale dell'esercizio: «la legge addirittura[258], non un decreto ministeriale, ha imposto che la quota premiale del fondo di finanziamento ordinario che deve arrivare progressivamente al 30% del totale, sia distribuita per almeno il 70% sulla base dei risultati della VQR. Quello che è stato fatto nel 2013-2014 andava già in questo senso. Quindi l'utilizzazione per la distribuzione dei fondi premiali è stata decisamente quella che ci aspettavamo anzi questa legge ha addirittura dato indicazioni più cogenti da questo punto di vista» (Intervista Benedetto).

Il terzo obiettivo, relativo all'allocazione delle risorse da parte di Atenei e strutture è forse quello maggiormente sentito e discusso nelle interviste. Nonostante non siano ancora chiari i modi in cui le strutture stanno utilizzando i dati è evidente un ampio utilizzo da parte soprattutto degli Atenei: «la CRUI sta completando un'indagine presso l'università per capire come sono stati utilizzati

[258] Il riferimento è all'articolo 60, comma 01, del Decreto Legge 21 giugno 2013, n. 69 (convertito con modificazioni dalla legge 9 agosto 2013, n. 98) che è intervenuto sulle modalità di attribuzione della quota premiale del FFO.

i risultati, per capire come sono stati distribuiti fondi interni e posti di ruolo, e di nuovo quello che emerge è che c'è stata un'utilizzazione molto significativa da parte degli atenei» (Intervista Benedetto)[259]. In effetti: «l'attenzione delle Università in questo ultimo anno è aumentata enormemente, anche perché il ministero ha sempre avuto delle reti di protezione, attenuava l'effetto negativo della quota premiale, ma questa rete di protezione si sta allargando, quindi in futuro la previsione potrebbe essere anche di perdite secche, di molti milioni di euro, per gli Atenei che performano male. Quello che vediamo in giro è una altissima sensibilità degli organi di vertice degli Atenei, e in particolare dei Rettori, che si trasferisce anche alle strutture decentrate e ai dipartimenti. Questo ha un effetto positivo, ma ha anche delle controindicazioni molto forti nel senso che in parte nevrotizza il sistema, in quanto lo rende fortemente soggetto a incentivi forti è chiaro che rende qualunque passo un passo molto importante» (Intervista Bonaccorsi). A proposito di questo aspetto sono state evidenziate alcune criticità: «ci sono state delle utilizzazioni distorte della VQR. Qualche università pensa di utilizzarla o lo ha già fatto per valutare le singole persone. Questo non funziona. Noi lo abbiamo detto in tutti i modi possibili però le università sono autonome in questo e qualche rettore ha addirittura chiesto di conoscere i risultati individuali. Ripeto, è un'utilizzazione sbagliata, per molti motivi che ho scritto nel rapporto finale, però qualcuno lo ha fatto[260]» (Intervista Benedetto).

Si tratta di una questione centrale, proprio perché rappresenta un utilizzo improprio di dati prodotti ad altri fini. Nel rapporto finale dell'Agenzia era stato chiaramente puntualizzato che: «i risultati della VQR non possono e non devono essere utilizzati per valutare i singoli soggetti. I motivi sono molteplici, e qui ne citiamo alcuni rilevanti: la scelta dell'associazione prodotti-soggetti valutati, dettata dall'ottimizzazione del risultato di struttura e non del singolo soggetto, la richiesta di conferire solo tre prodotti di ricerca pubblicati in sette anni, che costituiscono in molti settori della scienza un'immagine della produzione complessiva dei singoli soggetti molto parziale, la non considerazione del contributo individuale al prodotto nel caso di presenza di coautori, e, infine, l'utilizzo di metodi di valutazione la cui validità dipende fortemente dalla dimensione del gruppo di ricerca cui sono applicati» (Anvur, 2013a, p. 9).

Nonostante questa puntualizzazione, la stessa Agenzia ha previsto l'uso dei risultati della VQR per la valutazione della qualità della ricerca svolta dai membri dei Collegi dei Docenti nell'ambito della valutazione dei corsi di dottorato di ricerca (indicatori R_{VQR} e X_{VQR}[261]). Inizialmente (Anvur,

[259] Ad esempio nella testimonianza del professor Pacchioni: «nella nostra università la VQR, cioè il risultato della VQR dipartimento per dipartimento, perché poi i dati aggregati sono stati tradotti in un ranking di dipartimenti, dipartimenti classificati con dipartimenti omogenei a livello nazionale, quindi giuristi con giuristi, sociologi con sociologi, chimici con chimici. Questa classificazione è usata dal nostro ateneo per distribuire i punti organico cioè di fatto le risorse umane, gli assegni di ricerca e quindi con un peso rilevante» (Intervista Pacchioni).

[260] Simili le parole del Presidente Fantoni: «alcuni rettori, e anche dipartimenti, hanno cercato di utilizzare questi dati non adeguatamente, quindi individualmente per scopi altri, di promozione eccetera […] Noi non abbiamo mandato questi dati, tranne che ai singoli individui, ogni singolo individuo aveva il proprio dato ma non gli altri […] Mentre diciamo che nella globalità va bene… individualmente non va bene. Anche su questo un uso improprio c'è stato, quanto uno sforzo da questo punto di vista, abbiamo avuto una forte pressione per far vedere gli atti per pubblicare tutto al quale ci siamo opposti. C'è stata una forte pressione da questo punto di vista» (Intervista Fantoni).

[261] Nella VQR l'indicatore R_{jr} è dato dal rapporto tra il voto medio attribuito ai prodotti attesi della struttura *i*-esima nell'Area *j*-esima e il voto medio ricevuto da tutti i prodotti dell'Area *j*-esima, mentre l'indicatore X_{jr} è

2014a) era previsto l'uso dei risultati della VQR 2004-2010, ma a seguito dei riscontri ricevuti da ricercatori, società scientifiche, atenei, CRUI e CUN si è deciso di modificare i criteri (Anvur, 2014b)[262], prevedendo l'utilizzo dei risultati della prossima VQR 2011-2014 (Anvur, 2014c).

Le critiche a questa scelta sembrano molto sentite e non si limitano ad aspetti come l'aggiornamento o la scelta dei prodotti. Ad esempio il presidente del GEV di Scienze Chimiche contestava con forza la possibilità di un simile utilizzo della VQR 2004-2010, sia in ragione delle scelte effettuate nel corso della valutazione sia in ragione delle sue premesse: «la cosa inaccettabile è l'utilizzo di queste cose sui piccoli numeri. Cioè noi abbiamo preso una decisione su questa cosa: di non usare questi risultati per il singolo ricercatore, che è una follia furiosa […] noi abbiamo sorteggiato i lavori da mandare in peer review, il 10%, facendo attenzione che fosse il 10% per ogni struttura ma non per ogni ricercatore. Se questo viene usato sui ricercatori se un ha sfortuna e gli sono andati tutti e tre in peer review? Viene trattato peggio degli altri. Allora uno deve stare attento a quali sono i numeri più piccoli su cui queste cose si applicano, che è il vero problema […] noi avevamo chiesto garanzie all'Anvur […] dopodiché questo è stato interpretato nel senso che non sarebbero stati resi pubblici i dati, ma questo non è sufficiente perché nei fatti sono stati utilizzati» (Intervista Barone).

Gli indicatori connessi ai risultati della VQR 2011-2014 saranno calcolati sul Collegio nella sua totalità: «l'uso dei risultati della VQR sarà limitato alla valutazione dell'aggregato (Collegio dei docenti) e mai dei singoli componenti, i cui valori contribuiranno unicamente alla valutazione dell'insieme» (Anvur, 2014c, p. 5); inoltre: «nella consapevolezza che l'attribuzione dei prodotti ai singoli soggetti sarà fatta dalle strutture con l'obiettivo di massimizzare il risultato per la struttura, a scapito in taluni casi della attribuzione ai singoli dei loro prodotti "migliori", verranno scelti per la valutazione del collegio i prodotti che hanno ottenuto la valutazione migliore, fra tutti quelli presentati alla VQR dalla struttura con un membro del collegio come coautore» (*ibidem*).

Le questioni della valutazione dei singoli e dell'uso dei risultati della VQR su unità differenti dai dipartimenti universitari o dalle unità di ricerca sono solo parte del problema, infatti generalmente i Collegi sono multidisciplinari. Un'altra limitazione presentata nel rapporto finale circa l'uso dei dati della VQR riguarda il confronto tra i risultati ottenuti in Aree diverse: «lo sconsigliano i parametri di giudizio e le metodologie diverse di valutazione delle comunità scientifiche all'interno di ciascuna area» (Anvur, 2013a, p. 7). Il calcolo degli indicatori R_{VQR} e X_{VQR} tiene conto di questa questione attraverso una normalizzazione: «per il calcolo degli indicatori la normalizzazione verrà fatta sia

dato dal rapporto tra la frazione di prodotti eccellenti della struttura nell'Area e la frazione di prodotti eccellenti dell'Area (Anvur, 2013a). Nel caso della valutazione dei dottorati questi indicatori sono calcolati non sulla struttura, ma sul collegio nella sua composizione completa, considerando per ciascun membro i migliori *n* prodotti di cui risulti coautore tra quelli valutati.

[262] Nel documento di commento ai riscontri si legge che le osservazioni principali ricevute circa gli indicatori VQR sono due: «la prima ne contesta l'attualità, essendo gli indicatori riferiti a pubblicazioni che avranno nel 2015, nel migliore dei casi, cinque anni di "anzianità", e non tengono inoltre conto degli assunti dopo il 2010. La seconda si riferisce alla impossibilità, per gli atenei che non possiedono UGOV come strumento per gestire l'elenco delle pubblicazioni, di riconoscere le tre migliori pubblicazioni di cui un docente o ricercatore è coautore. Riconoscendo la fondatezza di tali osservazioni, l'Anvur ha deciso di non prendere in considerazione la VQR 2004-2010 e di utilizzare, invece, i risultati della prossima VQR 2011-2014, che saranno disponibili nella seconda metà del 2016. Per ovviare alla seconda critica, il CINECA predisporrà procedure per il conferimento dei prodotti che consentano la riconoscibilità di tutti i coautori dei prodotti afferenti allo stesso ateneo» (Anvur, 2014b, p. 1-2).

utilizzando la media nazionale a livello di SSD sia la media nazionale a livello di area, e si sceglierà il risultato migliore per il Collegio» (Anvur, 2014c, p. 5).

Dal punto di vista dell'Agenzia la VQR: «è designata per le strutture, però nel momento in cui arriva fino agli SSD e ai dipartimenti non siamo tanto lontani dai numeri di un collegio di dottorato che va, come minimo, da 16 persone, ma quasi nessuno ne ha solo 16, la media credo che gira intorno a 35-40, che sono i numeri di un dipartimento piccolo o medio. Di nuovo è comunque un gruppo significativo. Noi abbiamo sempre calcolato indicatori medi, su… appunto in quel caso, su almeno una ventina di persone… quindi crediamo, tutto sommato, che sia abbastanza ragionevole. Di nuovo, si evita di fare una valutazione *ad hoc* per il dottorato, che sarebbe stata molto complicata» (Intervista Benedetto).

I correttivi apportati ai criteri per la valutazione della qualità della ricerca svolta dai membri dei Collegi dottorali potrebbero essere in grado di ovviare, almeno in parte, alle obiezioni già esposte, nondimeno potrebbero condurre a ulteriori pressioni a rendere noti agli organi di *governance* i risultati individuali della VQR al fine di effettuare scelte mirate al momento della scelta dei membri dei Collegi, con esiti non necessariamente prevedibili. Nel caso della VQR 2004-2010 era stata prevista la pubblicazione del database completo, depurato dai dati sensibili[263], ma ad oggi l'insieme dei dati in questione non è ancora stato reso pubblico proprio per tutelare i soggetti valutati ed evitare qualsiasi uso improprio dei dati relativi alle valutazioni dei singoli prodotti[264].

L'ultimo obiettivo della VQR era la valutazione della posizione del sistema nazionale della ricerca nel panorama scientifico internazionale, ed è stato realizzato tramite un'analisi bibliometrica che considerava fattori di *output* (pubblicazioni e citazioni) e fattori di *input* (numero di ricercatori e spesa in ricerca e sviluppo)[265] (Anvur, 2013c). La produzione scientifica, l'impatto della ricerca, la collaborazione scientifica (nazionale ed internazionale) del sistema nazionale sono state esaminate e confrontate con il panorama internazionale, insieme all'analisi della produttività e dell'eccellenza scientifica.

Dal punto di vista informativo tutti gli obiettivi della VQR 2004-2010 sono stati raggiunti, d'altra parte gli esercizi di valutazione non danno solo risultati, producono degli impatti: più o meno attesi e più o meno desiderati.

---

[263] «L'Anvur, per motivi di trasparenza e per mettere a disposizione della comunità scientifica non solo nazionale l'enorme mole di dati derivanti da quello che è il più vasto esercizio di valutazione mai tentato nel nostro paese, intende rendere pubblico il *database* della VQR dopo averlo depurato dei dati sensibili. La disponibilità dei dati della VQR consentirà di proporre e sperimentare nuovi indicatori bibliometrici e di approfondire, a partire da dati reali, il dibattito in corso sui vantaggi e gli svantaggi della valutazione bibliometrica e della *peer review* insieme a molti altri temi di interesse» (Anvur, 2013a, p. 10).

[264] Il sito dell'Agenzia riporta che: «l'articolo 12 (Trasparenza) del DM del 7 luglio 2011 recita: "Ai sensi dell'art. 6, comma 4, del decreto legislativo 5 giugno 1998, n. 204, sarà cura dell'Anvur diffondere i risultati della VQR 2004-2010, compresi i giudizi sulle singole pubblicazioni valutate, fermo restando il rispetto dell'anonimato degli esperti". Poiché però la "trasparenza" non deve collidere con il rispetto della privacy, l'esito delle singole valutazioni sarà unicamente inserito nella pagina personale di ciascun ricercatore autore del prodotto e degli eventuali co-autori afferenti alla stessa struttura. Sarà resa pubblica, invece, la valutazione aggregata» (Anvur: http://www.Anvur.org/index.php?option=com_content&view=article&id=117&Itemid=232&lang=it).

[265] Questa analisi, pur se focalizzata sul periodo 2004-2010, ha considerato tutte le informazioni disponibili sul periodo 1981-2010 e si è basata sulla classificazione disciplinare delle Aree CUN. Come nella valutazione dei prodotti anche qui l'analisi bibliometrica si è concentrata sulle Aree 1-9, l'Area 13 e parte dell'Area 11.

7.2 I possibili impatti

Esiste una letteratura ampia e articolata sulla produzione scientifica, riconducibile in massima parte alla sociologia della scienza e all'ambito scientometrico. In questa letteratura trova spazio anche l'impatto dei diversi sistemi di valutazione della ricerca sulla produzione scientifica, seppure con studi spesso limitati a uno o più dipartimenti o università, e solo raramente di respiro nazionale.

I sistemi di valutazione della ricerca vengono implementati per incentivare e controllare la produzione individuale o delle strutture dedicate, tuttavia la letteratura su quale sia il loro impatto, e tramite quali processi questo si realizzi, è molto ridotta (Osuna *et al.* 2010) e, andrebbe aggiunto, per lo più critica. Gli studi comparativi sono pochissimi, alcuni paesi infatti hanno sistemi "forti", altri "deboli" (Whitley e Glaser, 2007), in alcuni di essi ad essere valutati sono gli individui, in altri le università oppure i dipartimenti (Geuna e Martin, 2003). In genere i risultati di questi studi risultano molto generali e poco significativi; in sostanza evidenziano come i sistemi di valutazione della ricerca incrementino la pressione a pubblicare sui ricercatori, rinforzando la cultura del *publish or perish* (Osuna *et al.* 2010). Geuna e Martin sottolineano come la valutazione della ricerca, soprattutto se legata al finanziamento della stessa, porta ad una sorta di inflazione delle pubblicazioni, senza necessariamente migliorarne la qualità (Geuna e Martin, 2003), producendo quello che è stato definito *salami slicing effect*, cioè la pubblicazione di batterie di articoli invece che un'unica pubblicazione, magari sotto forma di monografia, in grado di rendere conto in maniera completa e sistematica dei risultati della stessa ricerca (Liefner, 2003). E' stato inoltre evidenziato come gli aspetti attesi dei sistemi di valutazione della ricerca sulla produzione scientifica dovrebbero tenere conto anche di fattori intervenienti, come il contesto del finanziamento, il campo scientifico e lo stadio della carriera dei singoli ricercatori (Whitley, 2007; Osuna *et al.* 2010).

La maggior parte degli studi sull'impatto della valutazione della ricerca è stato svolto in paesi in cui il sistema è "forte", in cui cioè i risultati della valutazione influiscono direttamente sul finanziamento (Osuna *et al.* 2010). Ad esempio nel Regno Unito, nel corso dei primi anni '90, il RAE aumentò la pressione a pubblicare per gli accademici e, secondo diversi autori, migliorò effettivamente la qualità della ricerca universitaria (Georghiou *et al.* 2000; Talib, 2001). I ricercatori inglesi hanno modificato il proprio comportamento in relazione alle pubblicazioni in due modi: «mirando a riviste con un elevato *impact factor* ed incrementando la presentazione di articoli prima di una scadenza RAE» (Georghiou *et al.* 2000, p. 46).

Un'altra evidenza interessante è che al cambiamento dei criteri di valutazione del RAE sia corrisposta una evoluzione ulteriore dei modelli di pubblicazione (Moed, 2008): prima del RAE del 1992 i ricercatori britannici avevano aumentato le pubblicazioni, il RAE del 1996 ha prodotto un aumento del fattore di impatto medio delle riviste di pubblicazione (nei criteri si era verificato uno spostamento dell'accento dalla quantità alla qualità), prima del RAE del 2001 è aumentato il fenomeno del co-autoraggio interno al Regno Unito, sebbene non fosse aumentata la produttività (evidenza interpretata come una risposta alle modifiche apportate ai criteri di valutazione, mirata alla riduzione del numero di ricercatori inattivi).

Le evidenze disponibili circa gli impatti del RAE nel Regno Unito chiariscono quanto l'implementazione di sistemi di valutazione della ricerca possa effettivamente influire sulla produzione scientifica. Le conseguenze principali sembrano collegate principalmente al tipo di pubblicazione e alle riviste, eppure sono differenti da paese a paese, soprattutto in ragione delle

diverse caratteristiche dei sistemi di valutazione della ricerca (a titolo di esempio si vedano: Westerheijden, 1997; Gläser *et al.* 2002; Butler, 2003a; Butler, 2003b; Jimenez-Contreas *et al.* 2003).

E' stato sottolineato come «la più grande lezione derivante dal RAE del Regno Unito […] è un equivalente per gli scienziati sociali del principio di indeterminatezza: questi esercizi influenzano il comportamento degli osservati, spesso in modi non previsti. Qualsiasi cosa sia misurata viene messa in evidenza, probabilmente a spese di ciò che non lo è[266]» (Macilwain, 2010). Bornmann ha denominato «mimetismo» questo fenomeno, infatti: «come alcuni animali che cercano di sfuggire ai loro predatori o cacciare le proprie prede (ad esempio attraverso l'assimilazione del loro aspetto all'ambiente), gli scienziati applicano strategie che dovrebbero consentire loro di raggiungere l'accountability bibliometrica (Rodriguez-Ruiz, 2009) e di assicurare fondi alle proprie ricerche[267]» (Bornmann, 2011a, p. 174). Lo stesso Bornmann cita tra i comportamenti adattivi adottati dai ricercatori, oltre al salami slicing effect e alla spinta a pubblicare su riviste indicizzate, la scelta di tematiche di ricerca *mainstream* per aumentare le probabilità di accettazione dei loro articoli da parte delle riviste, la scelta di ricerche a breve termine per poterne pubblicare più velocemente i risultati.

E' chiaro che: «i cambiamenti nel comportamento degli scienziati attraverso questo sistema sono in effetti attesi e voluti[268]» (Schneider 2009), nel senso che la valutazione mira anche, forse soprattutto, al miglioramento delle pratiche in uso e dei loro risultati; è altresì evidente che diverse possibili conseguenze sono inattese e a volte indesiderate.

Stando alla ricognizione della letteratura sull'impatto dei sistemi di valutazione della ricerca è possibile avanzare una serie di ipotesi specifiche sui mutamenti nelle caratteristiche della produzione scientifica nel sistema della ricerca italiano. In generale è possibile attendersi un aumento della quota degli articoli su rivista sul totale dei prodotti, in secondo luogo è ipotizzabile un cambiamento nelle loro caratteristiche. Ad esempio nelle aree non bibliometriche (come l'Area di Scienze Politiche e Sociali) alla luce del dibattito sulla classificazione delle riviste e di quello sugli esiti delle valutazioni dovrebbe risultare identificabile una tendenza a pubblicare maggiormente su riviste dotate di sistemi di referaggio cieco, magari a diffusione internazionale, mentre nelle aree bibliometriche (come l'Area delle Scienze Chimiche) la tendenza dovrebbe risultare diretta alla pubblicazione di articoli in riviste con elevati indici di impatto, ma anche non troppo specialistiche e settoriali quanto a contenuti.

Nel corso delle interviste queste questioni hanno trovato spazio, seppure in modo molto diverso da intervistato a intervistato, è dunque possibile esporre i punti di vista di alcuni tra i testimoni a proposito degli impatti attesi e i possibili impatti indesiderati della VQR.

7.2.1 Gli impatti attesi

Il primo impatto della VQR, nonché in generale dei diversi esercizi condotti negli ultimi anni dall'Anvur, è indubbiamente il rafforzamento della cultura della valutazione nel mondo universitario: ciò che per lunghi anni era apparso come una pratica amministrativo/burocratica priva di impatti sulla vita accademica (*cfr.* Capitolo 1) è ora percepito come un elemento essenziale ai fini

[266] Traduzione dall'originale in lingua inglese.
[267] Traduzione dall'originale in lingua inglese.
[268] Traduzione dall'originale in lingua inglese.

dell'ottenimento di risorse economiche, umane e reputazionali. Si tratta di un impatto atteso e desiderato, riconosciuto sia dai componenti del Consiglio Direttivo Anvur sia dai membri dei GEV intervistati: «la cultura della valutazione ormai è stata assunta dai colleghi e da tutta l'università e quindi c'è un'attenzione molto grande a come viene fatta e ai risultati» (Intervista Benedetto).

Già la realizzazione della VTR aveva costituito un passo in avanti verso l'assimilazione della valutazione da parte dell'università italiana (Reale, 2013), tuttavia la sostituzione formale del CIVR con l'istituzione dell'Anvur nello stesso anno della pubblicazione dei risultati dell'esercizio ne ha smorzato gli effetti. La VQR invece è stata indetta e portata a termine in un clima di sviluppo e rafforzamento della valutazione a livello istituzionale e legislativo, da un'Agenzia con un mandato forte, e in un quadro di austerità economica che ha accentuato l'importanza dell'allocazione delle risorse e del reclutamento del personale.

Con specifico riferimento alla valutazione della ricerca, produttività e qualità sono i nodi centrali: «credo che sia divenuta consapevolezza diffusa innanzitutto che la ricerca è un elemento imprescindibile nella vita di un docente universitario, tanto più un ricercatore che, è nella parola stessa, che quindi in qualche modo deve porsi rispetto al problema di produrre. L'altra cosa che credo sia abbastanza entrata è l'idea che non basta produrre bisogna anche che questa cosa venga apprezzata» (Intervista Torrini). Entrambi questi aspetti sono ben presenti tanto agli intervistati dell'Area di Scienze Chimiche quanto a quelli dell'Area di Scienze Politiche e Sociali.

Circa la produttività, la spinta della VQR è piuttosto ridotta, dato che su una finestra di sette anni si richiedevano tre pubblicazioni per soggetto, nondimeno: «per la prima volta è stato chiaro a tutti che non pubblicare è un problema ed è un problema grosso. […] Se non pubblichi il tuo status crolla a quello di inattivo. E questo secondo me è la "vera" rivoluzione di questo esercizio» (Intervista Torsi).

L'accento sulla qualità delle pubblicazioni è, ovviamente, molto più forte e spinge verso la stessa direzione sia le scienze dure che le scienze umane e sociali anche se con connotazioni diverse. In discipline dove standard bibliometrici di qualità delle pubblicazioni sono disponibili e condivisi, si evidenzia «una maggiore attenzione alle riviste su cui uno pubblica, se mai fare qualche lavoro in meno, ma sulle riviste migliori, per dirne una, o fare più attenzione a che cosa uno sceglie» (Intervista Barone); per contro nelle discipline in cui i criteri di valutazione delle pubblicazioni sono meno standardizzati «si è capito, primo: che un qualche criterio di " serietà scientifica", "qualità scientifica", va comunque rispettato […] si è avviato un processo, un processo che a mio parere ci sta portando a standard più condivisi internazionalmente» (Intervista Colozzi).

L'importanza di questo aspetto viene percepita anche al di là dell'esercizio di valutazione, ad esempio in relazione alle carriere, sia in Area 3: «per la prima volta stiamo dicendo a chiare lettere che la carriera universitaria si fa grazie alle pubblicazioni e che devono essere non pubblicazioni qualunque, devono essere anche prodotti di qualità» (Intervista Torsi); sia in Area 14: «la gente ha cambiato la testa, cioè comincia a pensare nei termini della valutazione, quindi soprattutto dell'internazionalizzazione, i giovani lo sentono moltissimo. I giovani avvertono il peso della valutazione del prodotto e quindi della ricerca della rivista che vale qualcosa, dell'indicizzazione, della citazione e anche della produttività» (Intervista Bazzicalupo).

L'elemento generazionale è chiamato in causa dai membri di entrambi i GEV e dell'Agenzia in riferimento tanto ai possibili impatti in termini di produttività della ricerca, quanto agli impatti più o meno desiderabili sulle caratteristiche della produzione scientifica ed in particolare riguardo

l'internazionalizzazione. Uno degli impatti desiderati della VQR era indirizzare la produzione scientifica nazionale verso riviste di elevato profilo: «l'aver utilizzato indicatori come l'indice citazionale e l'*impact factor* o equivalente dà delle indicazioni positive ai giovani soprattutto nel senso di indirizzare le pubblicazioni dove è più difficile pubblicare, ma dove ha più senso» (Intervista Benedetto). A questo proposito, chiaramente, la costruzione di standard nelle scienze umane e sociali è meno semplice e indolore dell'uso di standard pre-esistenti nell'ambito delle scienze dure.

Dal punto di vista dei valutatori, la VQR e il dibattito che ne è seguito hanno avuto un impatto forte: «ormai è entrato il fatto che non tutte le riviste sono equivalenti […] che c'è quindi un problema di comitato editoriale, di *board*, di consulenti internazionali. Tutto questo si è mosso, in realtà, in questi anni. Si è mosso e si è mosso positivamente» (Intervista Colozzi). In relazione alla rilevanza delle riviste ed in particolare delle loro classificazioni in classi di merito, più della VQR, è l'Abilitazione Scientifica Nazionale (ASN) a risultare determinante: «la classificazione delle riviste nelle scienze umane e sociali, che non è stata utilizzata per la valutazione, e che è invece un'azione di fatto imposta dall'Abilitazione Scientifica Nazionale, di nuovo con luci e ombre, io credo che alla fine qualche indicazione utile la dia. Perché se è vero che nelle aree delle scienze dure e della vita le riviste non sono tutte uguali, io credo che la stessa cosa valga per le scienze umane e sociali. Quindi ciascuno dei colleghi in quelle aree sa benissimo che ci sono riviste migliori di altre. Certamente il processo è molto più complicato perché non ci sono indicazioni oggettive. O meglio non ci sono indicazioni oggettive legate appunto alle citazioni e la cultura citazionale in quelle aree è del tutto diversa. Quindi bisogna creare dei criteri che mettano insieme reputazione e anche elementi oggettivi. […] Forse varrebbe la pena di cercare di affinare un pochino questa classificazione. Anche se la sua utilizzazione per valutare il singolo articolo continuo a ritenerla difficile se non sbagliata» (Intervista Benedetto).

Un'ulteriore conseguenza dell'implementazione dell'attuale sistema di valutazione della ricerca è stato individuata dal Presidente del GEV di Area 14 nell'ampliamento della pratica del referaggio, dunque anche nell'avvicinamento all'individuazione di standard valutativi condivisi: «da noi la pratica del referaggio è ancora molto poco diffusa. Se lei pensa il referaggio delle riviste è diventato abituale negli ultimi tre-quattro anni, prima praticamente non esisteva se non in pochissime riviste, alcune delle quali lo dichiaravano e non lo facevano, tra l'altro, diciamo che in ogni caso era molto limitato. E' escluso ancora quasi totalmente dalle monografie. Quindi in realtà gli italiani che hanno fatto un lavoro di referaggio serio non sono poi così tanti, e soprattutto quelli che lo hanno fatto non hanno esperienza europea, non hanno esperienza internazionale, perché quei pochi che lo hanno fatto è perché facevano già da *referee* delle riviste ad *impact factor* che si rivolgevano a loro per prodotti magari di italiani o comunque su tematiche… venivano coinvolti dalle direzioni di queste riviste. Tranne questa piccolissima cerchia di persone che ha maturato una esperienza internazionale e quindi ha un idea di standard di qualità, per gli italiani questo standard probabilmente non c'è. Questo è dipeso quindi non solo dal fatto che le riviste… non solo le riviste, in generale la diffusione della peer review era ancora molto limitata, ma anche dal fatto che nell'unico caso di valutazione precedente, quella del CIVR, il numero di docenti coinvolti fu molto basso perché i criteri erano completamente diversi, cioè ogni dipartimento doveva presentare due lavori, per dipartimento, non tre per persona» (Intervista Colozzi)[269].

[269] Lo stralcio è già stato riportato nella nota 181 (p. 140) a sostegno dell'argomentazione, si è tuttavia ritenuto di un certo interesse riproporlo qui per la sua valenza interpretativa.

L'uso dell'internazionalizzazione come criterio di valutazione della qualità scientifica «è un messaggio chiaro alle comunità scientifiche. Insomma non è pensabile che le future carriere dei giovani siano costruite senza un riferimento internazionale, declinato nei modi in cui ogni comunità può farlo […] Non possiamo costruire carriere di giovani studiosi oggi che prescindano da un solido ancoraggio internazionale. Sappiamo che questo ha creato scompiglio, ha creato problematiche anche molto delicate, perché intere discipline hanno i primi livelli, cioè gli ordinari, che hanno un grado di internazionalizzazione spesso molto limitato. Nessuno immaginava che sarebbe stato accolto a braccia aperte, è chiaro che alcuni maestri ritengono di essere stati spodestati della loro autorevolezza, ma non è così. Bisogna affidarsi al fatto che i processi sono evolutivi e che, dal punto di vista della proiezione internazionale della ricerca italiana, dobbiamo proseguire» (Intervista Bonaccorsi).

La spinta verso l'internazionalizzazione, dunque, è di certo uno degli impatti attesi e desiderati della VQR, con sensibili differenze di ricezione e interpretazione tra le Aree: «non a caso quelle che non sono state, ad oggi, problematiche, cioè le Aree 1-9 hanno accolto la VQR direi senza critiche, perché corrisponde alla loro modalità comune, normale di fare scienza. Ci sono alcune opinioni contrarie, minoritarie, rispettabili […] però fondamentalmente c'è un isomorfismo tra il modo con cui le comunità organizzano il loro lavoro e il modo in cui avviene la valutazione. Nelle Aree non bibliometriche la faccenda è sicuramente più complicata» per questa ragione è necessario tentare di «capire la transizione verso l'internazionalizzazione di alcune aree disciplinari, che non è un fatto di moda, ma è un fatto profondo, di apertura delle proprie pratiche scientifiche, dei propri criteri, a una competizione e a una conversazione internazionale. Cosa che ha una componente generazionale, in cui alcuni riescono e altri no, c'è una componente di equilibri di potere, di risposte: premia chi è più dinamico in queste comunità e penalizza chi ha invece maggiormente potere o autorità nazionale» (Intervista Bonaccorsi).

In effetti nelle interviste ai membri del GEV di Scienze Chimiche la questione dell'internazionalizzazione è stata trattata solo marginalmente, la sua centralità è data per scontata, mentre tutti i membri del GEV di Scienze Politiche e Sociali hanno segnalato questo aspetto, spesso facendo anche riferimento alle questioni generazionali e alle altre dinamiche citate da Bonaccorsi.

L'opinione del Presidente del GEV 14 sui possibili impatti della VQR sull'internazionalizzazione, nonostante le sue perplessità circa la formulazione del criterio, è positiva: «è passato il fatto di tentare di entrare nel mercato internazionale della ricerca, quindi cominciare a scrivere in inglese, cercare di pubblicare, qualcosa almeno, delle proprie ricerche su riviste internazionali, magari non subito sull'*American Sociological Review*, però con l'*impact factor*, eccetera. Questo è un impatto che di per sé la VQR ha prodotto, io credo ci sia stato, oggi tutti ne parlano, io vedo anche i più giovani cominciano a pensare in questi termini. Come dire, pensano alla propria carriera in questi termini» (Intervista Colozzi). E' positivo anche il bilancio del professor Cipriani: «noi, anche per una ragione di barriere linguistiche, in molti casi siamo rimasti provinciali. Non ho difficoltà a dire che molti professori ordinari della generazione precedente alla mia, non erano mai andati a un convegno internazionale, oggi questo non esiste […] certamente l'internazionalizzazione andava spinta, andava favorita, enfatizzata e questo è un dato positivo», nonostante alcuni rischi: «c'è il risvolto della medaglia, che è una tendenza a pubblicare comunque in una qualunque rivista che abbia un carattere internazionale, perché comunque questo fa agio rispetto a quelle che sono le valutazioni di

un'Abilitazione Scientifica Nazionale o di altro genere, naturalmente ivi comprese poi le valutazioni della VQR» (Intervista Cipriani).

L'effettivo impatto della VQR sulle modalità di comunicazione dei risultati della ricerca è ancora lontano dal poter essere valutato, ma ci si può attendere un cambiamento significativo: «questi sono risultati che si vedranno più tardi[270]. Certamente se uno va a vedere quello che le università stanno facendo per distribuire all'interno le risorse si vede chiaramente che si è passati da una fase in cui contavano le pubblicazioni, a una fase in cui introducono elementi qualitativi molto simili alla VQR. Quindi indubbiamente ci sarà un cambiamento nelle attitudini e nelle scelte pubblicistiche delle persone» (Intervista Benedetto).

7.2.2 I possibili impatti non desiderati

A distanza di soli due anni dalla pubblicazione dei risultati della VQR non è possibile valutare i suoi impatti reali. Un discorso analogo potrebbe valere, a maggior ragione, circa i possibili impatti inattesi, ed eventualmente non desiderabili.

Ciò nonostante è possibile rilevare le impressioni degli esperti valutatori e dei membri del Consiglio su quelli che potrebbero essere gli effetti indesiderabili della VQR e, soprattutto, degli usi impropri dei suoi risultati.

Gran parte degli effetti indesiderati della valutazione della ricerca è connessa all'adozione da parte di ricercatori di comportamenti che ledono la normale etica utilizzata nella comunicazione scientifica. E' stato acutamente osservato da Bornmann (2011a) che in questi casi il comportamento degli scienziati si configura, mertonianamente, come un comportamento anomico. Si tratta in effetti dell'adozione di comportamenti non istituzionalizzati per il raggiungimento di mete culturali imposte dalle strutture sociali (Merton, 1938). Nel caso della valutazione della ricerca si tratta di canoni già presenti nelle comunità scientifiche, che possono però assumere un'importanza crescente nel momento in cui i criteri di produttività e qualità stabiliti da Agenzia e Ministero diventano la base per l'allocazione delle risorse. In questo caso il raggiungimento delle mete (produttività, impatto) potrebbe superare in importanza il rispetto dei mezzi approvati per raggiungerle, dando luogo a comportamenti non conformi all'etica scientifica (pubblicazioni ridondanti; frammentazione della rendicontazione dei risultati; scambi nell'attribuzione di pubblicazioni; manipolazione degli indicatori citazionali).

E' possibile riportare alcuni esempi di comportamenti di questo genere citati dagli esperti valutatori nel corso delle interviste, curiosamente quasi tutti riferiti alla produttività, che pure è uno degli aspetti meno stressati dalla VQR, ma pure presenti in altri contesti valutativi, come la già citata ASN. Ad esempio: «gli aspetti negativi sono banalmente che uno mette sui lavori il nome di gente che non ha fatto niente, perché di inattivi si parla, lo dico brutalmente, e questa non è una cosa così favorevole. In più questo ha avuto un impatto sul mandare in pensione i ricercatori […] Dopodiché, in parte, questo ha un effetto positivo perché invita tutti a diventare più attivi; in parte, siccome è stato

[270] Anche il Presidente Anvur ha evidenziato che bisognerà attendere: «[l'impatto della VQR] non credo riusciremo a vederlo tanto bene nemmeno con il prossimo esercizio, è ancora un po' presto […] Facciamo funzionare questo sistema per un po' di anni, facciamo uno o due esercizi e valutiamolo dopo» (Intervista Fantoni).

estremizzato, ovviamente è diventato anche eccessivamente penalizzante e marginalizzante» (Intervista Barone).

Sempre in relazione alla produttività viene citato anche il *salami slicing effect*, soprattutto dagli esperti di Area 14: «ci si butta a pubblicare più sulle riviste e non invece a dedicarsi a opere monografiche ponderate, magari anche poderose, preparate con un adeguato impegno anche in termini di tempo, perché il tempo è anche importante nell'attività di ricerca. Questo è il primo effetto deleterio: sempre più ormai ci si orienta a pubblicare su una rivista, anziché a fare una pubblicazione in termini di volumi» (Intervista Cipriani). La scelta dell'articolo su rivista come principale canale di comunicazione sarebbe un effetto né del tutto inatteso né del tutto indesiderato[271]: è la frammentazione della comunicazione scientifica a rappresentare un rischio, un impatto non desiderabile, dei processi di valutazione.

Da parte dell'Agenzia l'impressione è che questo rischio non sia poco esteso, in ragione del numero ridotto di prodotti che viene richiesto a ciascun ricercatore per la valutazione: «ci sono degli effetti opportunistici, ci sono persone che si scambiano i lavori l'uno con l'altro per aumentare il numero dei prodotti, ci sono… però sono ancora fenomeni aneddotici per quello che si può osservare» (Intervista Bonaccorsi).

Non va sottovalutata neppure la possibilità che eventuali cambiamenti nella scelta dei canali comunicativi influenzino la scelta delle tematiche di ricerca. I rischi messi in luce dalla letteratura sono due: la scelta di tematiche e approcci *mainstream*, al fine di aumentare le possibilità di pubblicazione per gli articoli, e la scelta di ricerche a breve termine, al fine di poterne pubblicare velocemente i risultati. Quest'ultiima problematica è nota, e tempo fa ha ottenuto una certa risonanza anche al di fuori del mondo accademico grazie a una rassegna di studi a lunghissimo termine apparsa su *Nature* (Owens, 2013). E' vero che il numero di prodotti richiesto dall'esercizio di valutazione è esiguo, nondimeno vi sono progetti che richiedono anni (a volte decenni) di studio, l'esempio classico è quello degli esperimenti a lungo termine e degli studi longitudinali, ma vale lo stesso discorso per quel genere di studi storici, filosofici, letterari, ecc. che implicano una lunga fase di apprendimento e documentazione precedente e necessaria alla fase di analisi e interpretazione. Questa eventualità è stata evidenziata in qualche maniera dai membri del GEV di Area 14, sempre in relazione alla scelta di pubblicare articoli piuttosto che opere monografiche, ad esempio: «ora che si sa che cosa è valutato tutto viene costruito in vista della valutazione, non in vista della ricerca: in vista della valutazione. Non so: un tempo uno avrebbe scritto un libro, mettendoci un sacco di anni, perché quel libro potesse essere un po'discusso da tutti, ora non vale più la pena» (Intervista Bazzicalupo).

[271] E' chiaramente espresso dal Presidente GEV di Area 14 in uno stralcio già riportato nella nota 187 (p. 142), ma che vale la pena rileggere anche in questo contesto: «la grande idea, che secondo me rappresenta un punto di non ritorno, è il fatto di aver cominciato a far capire che il veicolo di comunicazione della ricerca scientifica non è solo la monografia, o non è prevalentemente la monografia. Anzi, che sarebbe bene spostarsi ancora di più dalle monografie agli articoli su rivista, ai saggi su rivista. Anche per il modo con cui adesso si fa ricerca, perché vista l'assoluta mancanza di finanziamenti […] Cambiare la comunicazione della ricerca, il modo di comunicare il lavoro di ricerca, per cui la monografia è sicuramente importante, la monografia ha un suo significato, nel nostro settore non credo sia eliminabile. Non possiamo arrivare a Medicina per cui i libri, gli unici libri che si scrivono sono i manuali. Non credo che si possa arrivare lì. Noi continueremo a scrivere delle monografie a carattere di ricerca, però non sono il veicolo di eccellenza della comunicazione, il veicolo deve diventare un altro» (Intervista Colozzi).

L'altra serie di rischi, connessi alle prospettive di studio, alle ricerche interdisciplinari, all'impatto sulla didattica è stata citata esclusivamente dal professor Bonaccorsi: «il rischio di ridurre il pluralismo, di rafforzare le aree del *mainstream* e delle ortodossie scientifiche, il rischio di fare solo ricerca e niente didattica, il rischio di creare dei monopoli, che sono concettualmente anche legittimi però vanno valutati empiricamente […] riconosco una serie di argomenti, cerco di stare molto attento in quella direzione, ma non vedo una scala di fenomeni gravi di distorsione» (Intervista Bonaccorsi).

Conclusioni

In sintesi il quadro che emerge dalle interviste evidenzia le differenze nelle pratiche di comunicazione scientifica e nelle abitudini valutative proprie delle comunità di riferimento delle due Aree selezionate come casi studio. Nell'Area delle Scienze Chimiche si evidenzia soprattutto un rafforzamento di standard valutativi già condivisi dalle comunità e i timori sono riferiti alla possibilità che i singoli o le strutture assumano comportamenti non virtuosi per il raggiungimento degli obiettivi. Nell'Area delle Scienze Politiche e Sociali, invece, l'accento è sul cambiamento negli standard valutativi dei singoli, non sempre valutato positivamente. La spinta all'internazionalizzazione è il nodo cruciale in quest'Area, insieme alla nuova centralità assunta dalle riviste e dalla revisione tra pari. Alcuni temono una frammentazione dei risultati scientifici come conseguenza della VQR (in cui le valutazioni delle monografie non sono risultate incoraggianti) e dell'ASN (in cui gli articoli su rivista rivestono un importanza fondamentale), altri ritengono il passaggio dalla monografia all'articolo un passaggio positivo, purchè non finisca per azzerare la produzione di opere monografiche.

Nonostante le differenze si notano alcune similarità nella messa in luce delle conseguenze della VQR sull'attitudine mentale dei ricercatori nei confronti della valutazione, della produttività e della selezione delle riviste su cui pubblicare. Una ulteriore similarità è nel forte riferimento generazionale nel momento in cui questi cambiamenti vengono ipotizzati o esposti.

Complessivamente gli impatti inattesi sono riferiti esclusivamente alle modalità della comunicazione scientifica e all'impatto mediatico, superiore alle aspettative dell'Agenzia soprattutto in relazione alle classifiche degli Atenei. Gli impatti che vengono giudicati negativamente, o comunque ritenuti non desiderabili, sono legati nuovamente soprattutto all'effetto mediatico delle classifiche e alla comunicazione scientifica, ma anche all'utilizzo improprio dei dati da parte delle strutture per la governance o l'allocazione delle risorse.

Conclusioni

Il concetto di qualità non è mai semplice da definire e rilevare perché è intrinsecamente relativo. Nella sua definizione è necessario il riferimento agli aspetti rilevanti (i criteri) e alla misura in cui gli oggetti cui il concetto deve essere riferito devono rispondere a ciascuno di questi aspetti (gli standard). La necessità di comparare gli oggetti tramite il riferimento a criteri, soprattutto in assenza di veri e propri standard, fa della *valutazione*, cioè dell'espressione di giudizi, l'unico mezzo per la rilevazione della qualità. Valutare la qualità dei prodotti della ricerca è un obiettivo particolarmente arduo: la produzione scientifica può assumere una enorme quantità di forme diverse e le caratteristiche che possono determinarne e/o denotarne la qualità sono altrettanto varie, di conseguenza l'individuazione e la definizione dei criteri di valutazione sono passaggi molto delicati.

La definizione della qualità della ricerca alla base della VQR 2004-2010 era eccessivamente vaga e ambigua rispetto allo scopo cui era preposta. I criteri devono essere individuati a partire dagli obiettivi dell'esercizio, possono essere più o meno condivisi dalle singole comunità scientifiche, ma devono essere definiti con chiarezza per poter essere tradotti in procedure operative ed esiti valutativi affidabili. Gli elementi di vaghezza e ambiguità delle definizioni semantiche dei criteri e delle classi di merito sono responsabili di gran parte delle problematiche metodologiche evidenziate, mentre la definizione quantitativa delle classi di merito, pur essendo più o meno condivisibile, è un criterio non affetto da ambiguità o vaghezza e risulta legato a molti tra i passaggi più trasparenti.

Il problema principale è che la definizione in questione, essendo parte integrante del decreto ministeriale che stabiliva l'esercizio di valutazione, non è stata discussa né avrebbe potuto essere modificata, neppure dall'Agenzia. Non si discute che al decisore politico spetti il ruolo centrale nella determinazione dei criteri di valutazione, ma l'Agenzia avrebbe dovuto svolgere un ruolo altrettanto fondamentale, controllando ed eventualmente discutendo la traducibilità in termini empirici delle definizioni ministeriali. Inoltre nell'ottica di una valutazione partecipata sarebbe stata opportuna una fase di condivisione e negoziazione con gli attori coinvolti, che avrebbe potuto dar luogo non solo a una maggiore adesione delle comunità scientifiche agli obiettivi dell'esercizio, ma anche a un largo controllo, se non metodologico per lo meno semantico, delle definizioni proposte. Una maggiore attenzione alla definizione dei criteri è, senza dubbio, il primo passo verso una Valutazione della Qualità della Ricerca più affidabile.

Le procedure di valutazione vere e proprie, per la VQR 2004-2010, sono state messe a punto dall'Agenzia, discusse nell'ambito della Conferenza dei Presidenti GEV e calibrate sulle esigenze delle varie Aree disciplinari dai GEV di riferimento. I GEV nel bando del 2011 sono definiti Gruppi di Esperti *della Valutazione*, ma nella loro selezione si è tenuto conto esclusivamente, oltre che della qualità scientifica e della continuità della loro produzione, della loro esperienza come *valutatori* (Anvur, 2013a). Ciò costituisce allo stesso tempo un punto di forza e un punto di debolezza. In quanto *esperti valutatori* i membri del GEV presentano le competenze necessarie alla selezione dei revisori, alla validazione delle loro valutazioni, alla conferma della valutazione finale di ciascun prodotto. Non necessariamente, però, a queste competenze si affiancano quelle di *esperti della valutazione*. In altri

termini anche il migliore tra i valutatori possibili potrebbe essere inconsapevole delle criticità di specifiche procedure di valutazione.

Questa problematica è particolarmente evidente in relazione alla valutazione diretta tramite analisi bibliometrica: un valutatore afferente al campo della Chimica, della Fisica, della Medicina, per quanto esperto nel proprio campo, non è necessariamente un esperto di scientometria Potrebbe dunque conoscere gli indicatori utilizzati, comprendere le procedure di classificazione, utilizzare in modo appropriato le informazioni che ne derivano, ma non essere pienamente consapevole dei limiti dei database esistenti o delle tematiche metodologiche connesse all'analisi biliometrica.

Allo stesso modo, in relazione alla peer review, gli esperti potrebbero sottovalutare le problematiche connesse agli strumenti utilizzati, in particolare alla formulazione della scheda di valutazione e alle procedure di sintesi, o quelle proprie della procedura: la selezione dei revisori, l'assegnazione dei prodotti, il carico di lavoro, l'uniformità delle scale di valutazione.

L'adeguatezza metodologica delle procedure va dunque affrontata a monte, a partire dalla definizione dei criteri, con la maggiore trasparenza possibile. L'esposizione delle criticità e dei punti di forza delle procedure potrebbe sì sollevare delle critiche, ma anche condurre al miglioramento di strumenti e algoritmi, a una maggiore condivisione delle pratiche valutative, a un clima di maggiore fiducia nei confronti dell'intero sistema. Non a caso la mancanza di trasparenza è stato uno dei nodi cruciali del dibattito sulla VQR.

La concretizzazione di una valutazione della ricerca più pubblica, ripetibile e controllabile passa attraverso una maggiore trasparenza nella rendicontazione delle procedure utilizzate, non solo in nome del principio di *accountability* e di una più generale etica scientifica, ma soprattutto perché solo attraverso il controllo e il contributo della comunità nazionale e internazionale si potrà ottenere il perfezionamento dei protocolli attualmente in uso.

Alla luce della letteratura e delle esperienze disponibili sembra necessario concludere che la procedura perfetta non esiste, né con riferimento alla peer review né con riferimento alla valutazione diretta tramite analisi bibliometrica. Ciò nonostante sono individuabili ampi margini di miglioramento rispetto alle procedure in uso: alcune proposte sono state avanzate, altre potrebbero essere immaginate apportando modifiche più sostanziali all'impianto generale dell'esercizio di valutazione, ad esempio rendendo possibile la valutazione di ciascun prodotto da parte di *panel* di esperti, con o senza il supporto informativo dell'analisi bibliometrica.

Un'ottima sintesi è la conclusione del cosiddetto Manfesto di Leiden: «le migliori decisioni si prendono combinando statistiche robuste con la sensibilità verso gli scopi e la natura della ricerca che viene valutata. C'è bisogno sia di evidenze quantitative che di evidenze qualitative, ognuna è obiettiva a suo modo. I processi decisionali sulla scienza devono essere basati su procedure di alta qualità, informate da dati di altissima qualità[272]» (Hicks *et al.* 2015, p. 431).

[272] Traduzione dall'originale in lingua inglese.

Ringraziamenti

E' stato un percorso lungo, per giunta non lineare né agevole, devo dunque molto a chi mi ha guidato o ha vegliato sui miei passi.

Ringrazio miei due tutor, Enzo Campelli ed Antonio Fasanella, non solo per la pazienza e la competenza con cui hanno seguito il mio lavoro, ma anche (forse soprattutto) per quanto mi hanno insegnato sulla complementarietà dei punti di vista, sulla necessità di analizzare criticamente il proprio lavoro, sull'importanza dell'argomentazione e della giustificazione nella rendicontazione della ricerca.

Ringrazio Paul Wouters, direttore del CWTS (*Centre for Science and Technology Studies*) dell'Università di Leiden (Paesi Bassi), per avermi dato modo di trascorrere due mesi presso l'Istituto, per approfondire le questioni metodologiche proprie della bibliometria avvalendomi del consiglio di ricercatori esperti nel campo. In particolare devo moltissimo a Thed van Leeuwen, che ha seguito con interesse ed impegno il mio lavoro, non solo durante la mia permanenza a Leiden, ma anche nelle fasi successive. Devo inoltre un ringraziamento a Martijn Visser, per aver indirizzato la mia attenzione alla classificazione dei documenti nei database e avermi fornito dati cui non avrei mai avuto accesso con le mie sole forze, oltre che per avermi consigliata circa l'analisi sulle *subject categories*; a Paul Wouters, Marc Luwel, Cornelis van Bochove, per i suggerimenti e le osservazioni ricevute, che pure, con riferimento al confronto con altri sistemi di valutazione, non è stato possibile concretizzare. Un ringraziamento va anche ad Henk Moed che nel corso della sua permanenza alla Sapienza come *visiting professor* nell'autunno 2014 mi ha concesso un breve ma fruttuoso scambio di opinioni.

Ringrazio, naturalmente, i membri dei GEV e dell'Anvur che si sono resi disponibili per le interviste focalizzate. Le informazioni sulle procedure e molte delle osservazioni avanzate personalmente dai testimoni hanno costituito spunti essenziali per l'avanzamento dell'analisi metodologica. Un ringraziamento particolare va a Giovanna Colizza e Alberto Anfossi, che pur non essendo tra i testimoni intervistati hanno fornito o integrato con estrema disponibilità e cortesia alcune informazioni di primaria importanza per l'analisi delle procedure.

Devo infine un ringraziamento a chi leggerà questo lavoro. Rubando le parole a Pascal mi scuso per la sua lunghezza, ma non c'è stato il tempo di renderlo più breve.

Glossario

AI	*Article Influence score*
Anvur	Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca
ASJC	*All Science Journal Classification*
CAR	Comitati di Area
CAT	Comitati di Ateneo
CIPE	Comitato interministeriale per la programmazione economica
CIVR	Comitato di Indirizzo per la Valutazione della Ricerca
CNVSU	Comitato Nazionale di Valutazione del Sistema Universitario
CRUI	Conferenza dei rettori italiani
CUN	Consiglio universitario nazionale
ERC	*European Research Council*
ETP	Equivalente a tempo pieno
EV	Esperto Valutatore
FFO	Fondo di Finanziamento Ordinario
FIRB	Fondo per gli investimenti della ricerca di base
GEV	Gruppo di Esperti della Valutazione
IF	*Impact factor*
IF5	*Impact factor* a cinque anni
IR	*Informed Review*
MIUR	Ministero dell'istruzione, dell'università e della ricerca
MURST	Ministero dell'Università e della Ricerca Scientifica e Tecnologica
OVSU	Osservatorio per la Valutazione del Sistema Universitario
PNR	Programma Nazionale di Ricerca
PRIN	Progetti di Rilevante Interesse Nazionale
SC	*Subject Category*
SNIP	*Source Normalized Impact per Paper*
SJR	*SCImago Journal Ranking*
SNR	Sistema Nazionale della Ricerca
SSD	Settori scientifico-disciplinari
VIU	Valutazione Istituzionale dell'Università
VPS	Valutazione della Produzione Scientifica
VQR	Valutazione Triennale della Ricerca
VTR	Valutazione della Qualità della Ricerca
WoS	*Web of Science*

Riferimenti bibliografici

AA. VV., 1990. *Enciclopedia delle scienze sociali*. Roma, Istituto dell'Enciclopedia Italiana.

AA.VV., 1999. *Thesaurus Italiano di Sociologia*. Firenze, IFNET.

Abramo, G., D'Angelo C.A., Di Costa, F., 2010. Citationsversus *impact factor* as proxy of quality: could the latter be preferable? *Scientometrics*, 84(3), pp. 821-833.

Abrizah, A., Zainab, A.N., Kiran, K., Raj, R. G., 2013. LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus. *Scientometrics*, 94(2), pp. 721-740.

Adam, D., 2002. Citation analysis: The counting house. *Nature*, *415*(6873), pp. 726-729.

Adriaanse, L., Rensleigh, C., 2013. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31(6), pp. 727-744.

Agnoli, M.S., 1992. La costruzione delle variabili tra rilevazione e analisi dei dati. in Marradi, A., Gasperoni, G., (a cura di) 1992. *Costruire il dato 2: vizi e virtù di alcune tecniche di raccolta delle informazioni*. Milano, Franco Angeli.

Agnoli, M.S., 1994. *Concetti e pratica nella ricerca sociale*. Milano, Franco Angeli.

Agnoli, M.S., 1999. Basic Issue: la «localizzazione dei problemi» nella metodologia lazarsfeldiana. in Campelli, E., Fasanella, A., Lombardo, C., (a cura di) 1999. *Paul Felix Lazarsfeld: un "Classico" Marginale.* Milano, Franco Angeli, pp. 189-214.

Agodi, M.C., 1999. Lazarsfeld e la «natura» della classificazione nelle scienze sociali. *Sociologia e ricercasociale*, 20(58/59), pp. 117-149.

Althouse, B.M., West, J.D., Bergstom C.T., Bergstom, T., 2009. Differences in *impact factor* across fields and over time. *Journal of the American Society for Information Science and Technology*, 60(1), pp. 27-34.

Amin, M., Mabe, M. 2000. *Impact factor*s: use and abuse. *Perspectives in Publishing*, 1, pp. 1-6.

Ammassari, P., 1995. *Saggi metodologici*. Milano, Franco Angeli.

Anvur, 2011. Valutazione della qualità della ricerca (VQR 2004-2010). Bando di partecipazione. http://www.Anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf.

Anvur, 2012. Documento di accompagnamento dei criteri VQR 2004-2010. (29.02.2012) http://www.Anvur.org/attachments/article/244/documento_accompagnamento_criteri.pdf

Anvur, 2013. Valutazione della Qualità della Ricerca 2004-2010 – Rapporto Finale. http://www.Anvur.org/ rapporto/main.php?page=intro 30.07.13.

Anvur, 2013a. Valutazione della Qualità della Ricerca 2004-2010 – Rapporto Finale. Parte Prima: Statistiche e risultati di compendio.

Anvur, 2013b. Valutazione della Qualità della Ricerca 2004-2010 – Rapporto Finale. Parte Seconda: La valutazione delle single strutture.

Anvur, 2013c. Valutazione della Qualità della Ricerca 2004-2010 – Rapporto Finale. Parte Terza: I confronti internazionali per le aree bibliometriche.

Anvur, 2013d. Valutazione della Qualità della Ricerca 2004-2010 – Rapporti di area. http://www.Anvur.org/ rapporto/main.php?page=intro 30.07.13.

Anvur, 2014. Criteri di assegnazione delle classi di merito nel caso di valutazioni peer review con valutazioni non coincidenti da parte dei *referee*. http://www.Anvur.org/attachments/article/244/Criteri%20di%20assegnazione%20della%20classe%20di%20merito%20.pdf 09.06.14.

Anvur, 2014a. La valutazione dei corsi di dottorato – Versione provvisoria http://www.Anvur.org/attachments/article/455/valutazione%20corsi%20dottorato_finale_clean.pdf

Anvur, 2014b. Commenti alle osservazioni sul documento Anvur "La valutazione dei corsi di dottorato". http://www.Anvur.org/attachments/article/455/CommentialleOsservazionisulDocValutazioneDottoratiFinale20141229.pdf

Anvur, 2014c. La valutazione dei corsi di dottorato – Versione definitiva http://www.Anvur.org/attachments/article/455/ValutazioneCorsiDottoratoFinale20141230.pdf

Archambault, É., Campbell, D., Gingras, Y., Lariviere, V., 2008. WOS vs. Scopus: On the reliability of scientometrics, *Book of Abstracts of the 10th International Conference on Science and technologyIndicators*, pp. 94-97.

Archambault, É., Campbell, D., Gingras, Y., Larivière, V., 2009. Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, 60(7), pp. 1320-1326.

Archambault, É., Vignola-Gagne, É., Côté, G., Lariviére, V., Gingrasb, Y., 2006. Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68(3), pp. 329-342.

Baggs, H.G., Broome, M.E., Dougherty, M.C., Freda, M.C., Kearney, M.H. 2008. Blinding in peer review: the preferences of reviewers for nursing journals. *Journal of Advanced Nursing*, 64(2), pp. 131-138.

Bailar, J.C., 1991. Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, 14(01), pp. 137-138.

Bakkalbasi, N., Bauer, K., Glover, J., Wang, L., 2006. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical digital libraries*, 3(1), p. 1-8.

Bar-Ilan, J., 2008. Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), pp. 257-271.

Baxt W.G., Waeckerle J.F., Berlin J.A., Callaham M.L., 1998. Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine,* 32(3), pp. 310-317.

Becker, S., Bryman, A., Sempik, J., 2006. *Defining 'Quality' in Social Policy Research: Views, Perceptions and a Framework for Discussion*. Lavenham, Social Policy Association.

Benos, D.J., Bashari, E., Chaves, J.M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S.,Polter, A., Qadri, Y.,Sarfare, S., Schultz, K.,Splittgerber, R., Stephenson, J., Tower, C. , Walton, R.G.,Zotov, A., 2007. The ups and downs of peer review. *Advances in physiology education*, 31(2), pp. 145-152.

Bentley, R., Blackburn, R., 1990. Changes in academic research performance over time: A study of institutional accumulative advantage. *Research in Higher Education*, 31(4), pp. 327-353.

Bichi, R., 2002. *L'intervista biografica. Una proposta metodologica*. Milano, Vita e pensiero.

Biolcati-Rinaldi, F., 2010. *Quali indicatori bibliometrici per le scienze sociali?*. Milano, Dipartimento di Studi Sociali e Politici, UNIMI; WorkingPaper 2.

Borgatta, E.F., Bohrnstedt, G.W., 1980. Level of measurement once over again. *Sociological Methods & Research*, 9(2), pp. 147-160.

Bornmann, L., 2008. Scientific Peer review: An Analysis of the Peer review Process from the Perspective of Sociology of Science Theories. *Human Architecture: Journal of the Sociology of Self-Knowledge*, 6(2), pp. 23-37.

Bornmann, L., 2011. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1), pp. 197-245.

Bornmann, L., 2011a. Mimicry in science?. *Scientometrics*, 86(1), pp. 173-177.

Bornmann, L., 2013. How to Analyze Percentile Citation Impact Data Meaningfully inBibliometrics: The Statistical Analysis of Distributions, Percentile Rank Classes, and Top-Cited Papers. *Journal of the American Society for Information Science and Technology*, 64(4), pp. 587-595.

Bornmann, L., Daniel, H.D., 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), pp. 45-80.

Bornmann, L., Daniel, H.D., 2009. The state of h index research. Is the h index the ideal way to measure research performance?. *EMBO reports*, 10(1), pp. 2-6.

Bornmann, L., Mutz, R., Daniel, H.D., 2007. Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, 1(3), pp. 226-238.

Bornmann, L., Mutz, R., Daniel, H.D., 2010. A reliability generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS One*, 5(12), e14331.

Boudon, R., Lazarsfeld, P.F., 1965. *Méthodes de la sociologie I Le vocabulaire des sciences sociales*. Paris, La Haye, Mouton& Co.; tr. it. 1969. *L'analisi empirica nelle scienze sociali: I Dai concetti agli indici empirici*. Bologna, Il Mulino.

Boudon, R., Lazarsfeld, P.F., 1966. *Méthodes de la sociologie II L'analyse empirique de la causalité*, Paris-La Haye, Mouton & Co.; tr. it. 1969 *L'analisi empirica nelle scienze sociali*: *IIL'analisi empirica della causalità,* Bologna, il Mulino,.

Bourke, P.F., Butler, L., Biglia, B., 1996. *Monitoring research in the periphery: Australia and the ISI indices*. Research Evaluation and Policy Project, Australian National University.

Boyack, K.W., Klavans, R., Börner, K., 2005. Mapping the Backbone of Science. *Scientometrics*, 64(3), pp. 351-374.

Bradford, S.C., 1934. Sources of information on specific subjects. *Engineering*, 137, pp. 85-86.

Braun, T., Glänzel, W., Schubert, A., 2006. A Hirsch-typeindex for journals. *Scientometrics*, 69(1), pp. 169-173.

Bruschi, A., 1996. *La competenza metodologica: Logiche e strategie nella ricerca sociale*. Roma, Nuova Italia Scientifica.

Brysbaert, M., Smyth, S., 2011. Self-enhancement in scientific research: the self-citation bias. *PsycologicaBelgica*, 51(2), pp. 129-137.

Burrows, R., 2012. Living with the h-index? Metric assemblages in the contemporary academy. *The Sociological Review*, 20(2), pp. 355-372.

Butler, L., 2003a. Modifying publication practices in response to founding formulas. *Research Evaluation*, 12(1), pp. 39-46.

Butler, L., 2003b. Explaining Australia's increased share of ISI publications - the effects of a founding formula based on publication counts. *Research Policy*, 31(1), pp. 143-155.

Callaham, M.L., Baxt, W.G., Waeckerle, J.F., Wears, R.L., 1998. Reliability of editors'subjective quality ratings of peer reviews of manuscripts. *JAMA: the Journal of the American Medical Association*, 280(3), pp. 229-231.

Campanario, J.M., 1998a. Peer review for journals as it stands today—Part 1. *Science Communication,* 19(3), pp. 181-211.

Campanario, J.M., 1998b. Peer review for journals as it stands today—Part 2. *Science Communication,* 19(4), pp. 277-306.

Campanario, J.M., 2015. Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal *impact factor*s. *Journal of Informetrics*, 9(2), pp. 398-407.

Campbell, D.T., Cook, T.D., 1979. *Quasi-experimentation: design and analysis for field settings*. Chicago, Rand McNally.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2), pp. 81-105.

Campbell, D.T., Stanley, J.C., 1966. *Experimental and quasi-experimental designs for research*. Boston, HoughtonMifflin Company; tr. it. 2004.*Disegni sperimentali e quasi-sperimentali per la ricerca*. Roma, Eucos.

Campelli, E., 1991. *Il metodo e il suo contrario: sul recupero della problematica del metodo in sociologia*. Milano, Franco Angeli.

Campelli, E., 1996. Metodi qualitativi e teoria sociale, in Cipolla, C., De Lillo, A., (a cura di), 1996. *Il sociologo e le sirene. La sfida dei metodiqualitativi*. Milano, Franco Angeli, pp. 17-36.

Campelli, E., 1999. *Da un luogo comune. Elementi di metodologia delle scienze sociali.* Roma, Carocci.

Campelli, E., Fasanella, A., Lombardo, C., (a cura di), 1999. *Paul Felix Lazarsfeld: un "Classico" Marginale*, numero monografico di *Sociologia e Ricerca Sociale*, 20(58/59). Milano, Franco Angeli.

Cannavò, L., 1995. Il primato della pragmatica. Il senso degli indicatori nella ricerca sociale. *Sociologia e ricerca sociale,* 16(47/48), pp. 7-26.

Cannavò, L., 1999. *Teoria e pratica degli indicatori nella ricerca sociale: teorie e problemi della misurazione sociale*. Milano, LED.

Cannavò, L., Basevi, M., 2003. Oltre Thurstone e Likert: la valutazione di atteggiamenti e motivazioni con la tecnica TLL. Roma, Euroma.

Cardano, M., Miceli, R., 1991. Il linguaggio delle variabili. Strumenti per la ricerca sociale. Torino, Rosenberg &Sellier.

Carnap, R., 1936. Testability and meaning. *Philosophy of science III*, pp.419-71; tr. it. 1971. *Analiticitàsignificanza, induzione*. Bologna, IlMulino.

Carnap, R., 1938. Logical foundations of the Unity of Science. in*International Encyclopedia of Unified Sciences, Vol. I*. Chicago, University of Chicago Press; tr. it. 1973. in AA. VV., *Neopositivismo e unitàdellascienza*. Milano, Bompiani.

Cartocci, R., 1984. Concetti e indicatori: il contributo della nuova retorica. *Sociologia e Ricerca sociale*, 13(5), pp. 69-98.

Ceci, S., Peters, D., 1982. Peer review-a study of reliability. *Change*, 14(6), pp. 44-48.

Ceci, S.J., Williams, W.M., 2011. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 108(8), pp. 3157-3162.

Centre for Quality Assurance and Evaluation of Higher Education (Denmark), Comite National d'Evaluation (France), 1998. *Evaluation of European higher education: a status report*. Copenhagen, European Commission/Centre for Quality Assurance.

Cho, M.K., Justice, A.C., Winker, M.A., Berlin, J. A., Waeckerle, J.F., Callaham, M.L., Rennie, D., 1998. Masking author identity in peer review. *JAMA: the Journal of the American Medical Association*, 280(3), pp. 243-245.

Chubin, D., Hackett, E. 1990. *Peerless science:* Peer review *and U.S. science policy*. Albany, State University of New York Press.

Cicchetti, D.V. 1980. Reliability of reviews for the American Psychologist: A biostatistical assessment of the data. *American Psychologist*, 35(3), pp. 300-305.

Cicchetti, D.V., 1991. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(01), pp. 119-135.

Cipolla, C., De Lillo, A. (a cura di), 1996. *Il sociologo e le sirene. La sfida dei metodiqualitativi*. Milano, Franco Angeli.

CIVR, 2003. *Linee guida per la valutazione della ricerca*. Roma, Mur. http://vtr2006.cineca.it/documenti/ linee_guida.pdf

CIVR, 2004. *Linee guida per i Comitati (panel) di Area per l'esercizio 2001-2003*. Roma, Murhttp://www.unicatt.it/nucleo_linee_guida_comitati.pdf

CIVR, 2006. *VTR 2001-2003. Risultati della valutazione dei panel di area*. Roma, Mur. http://vtr2006.cineca.it/ pubblicazioni/volume_completo.pdf

CIVR, 2006a. *VTR 2001-2003. Relazione finale*. Roma, Mur. http://vtr2006.cineca.it/php5/relazione_civr/output/ totale.pdf

CNSVU, 2002. *L'evoluzione del sistema di valutazione negli ultimi 10 anni*, «Atenei», 1.

Cole, J. R., 2000. The role of journals in the growth of scientific knowledge. inCronin, B., Atkins, H.B., (a cura di), 2000, *The web of knowledge. A festschrift in honor of Eugene Garfield*. Medford, Information Today, pp. 109-142.

Cole, S., Cole, J.R., Simon, G.A., 1981. Chance and consensus in peer review. *Science*, 214(4523), pp. 881-886.

Coleman, J. S., 1964. *Introduction to mathematical sociology*. London, Free Press Glencoe.

Colledge, L., de Moya-Anegón, F., Guerrero-Bote, V., López-Illescas, C., ElAisati, M.H., Moed, H., 2010. SJR and SNIP: two new journal metrics in Elsevier's Scopus. *Serials: The Journal for the Serials Community*, 23(3), pp. 215-221.

Cook, C., Heath, F., Thompson, R.L., Thompson, B., 2001. Score reliability in Web or internet-based surveys: unnumbered graphic rating scales versus Likert-type scales. *Educational and PsychologicalMeasurement*, 61(4), pp. 697-706.

Corbetta, P., 1999. *Metodologia e tecniche della ricerca sociale*. Bologna, Il Mulino.

Costas, R., Bordons, M. van Leeuwen, T., 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 American Society for Information Science and Technology,* 60(4), pp. 740-753.

Cozzens, S.E., 1989. What do citation count? The rhetoric-first model. *Scientometrics,* 15(5), pp. 437-444.

Crandall, R., 1978. Interrater agreement on manuscripts is not so bad!.*American Psychologist,* 33(6), pp. 623-624.

Crane, D., 1965. Scientists at major and minor universities: A study of productivity and recognition. *American sociological review*, 30(5), pp. 699-714.

Crane, D., 1967. The gatekeepers of science. Some factors affecting the selection of articles in scientific journals. *American Sociologist*, 2(4), pp. 195-201.

Cronin, B., 1981. The need for a theory of citing. *Journal of Documentation*, 37(1), pp. 16-24.

Cronin, B., Atkins, H.B., (a cura di), 2000, *The web of knowledge. A festschrift in honor of Eugene Garfield*. Medford, Information Today.

Davidoff, F., 1998. Masking, blinding, and peer review: the blind leading the blinded. *Annals of internalmedicine*, 128(1), pp. 66-68.

Demazière, D., Dubar, C., 1997. *Analyser les entretiens biografiques*. Paris, Nathan; tr.it. 2000.*Dentro le storie. Analizzare le interviste biografiche*. Milano, Cortina Raffaello Editore.

Di Ciaccio, A., Borra, S., 1996. *Introduzione alla statistica descrittiva*. Milano, McGraw-Hill.

Di Giammaria, L., 2009. *Realismo scientifico e disposizioni sociali: teoria sociologica e metodologiadella ricerca sociale nel dibattito tra realisti e costruttivisti*. Acireale-Roma, Bonanno.

Dorta-Gonzalez, P., Dorta-González, M. I. 2013. Impact maturity times and citation time windows: The 2-year maximum journal *impact factor*. *Journal of Informetrics*, 7(3), pp. 593-602.

Durrant, G.B., 2009. A typology of research methods within the social sciences. *eSocialSciencesWorking Papers*, id:2003.

Eckberg, D. L., 1991. When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences*, 14(01), pp. 145-146.

Egghe, L., 2005. *Power laws in the information production process: LotkaianInformetrics*. Oxford, Elsevier.

Egghe, L., 2009. Mathematical derivation of the *impact factor* distribution. *Journal of Informetrics*, 3(4), pp. 290-295.

Egghe, L., Waltman, L., 2011. Relation between the shape of the size-frequency distribution and the shape of a rank-frequency distribution. *Information Processing and Management*, 47(2), pp. 238-245.

Elkins, M.R., Maher, C.G., Herbert, R.D., Mosesley, A.M., Sherrington, C., 2010. Correlation between the Journal *Impact factor* and three other citation indices. *Scientometrics*, 81(1), pp. 81-93.

EU (European Commission: Expert Group on Assessment of University-Based Research), 2010. *Assessing Europe's University Based Research*, European Commission Directorate General for Research. Luxembourg, Publications Office of the European Union.

Evans, J.T., Nadjari, H.I. Burchell, S.A., 1990. Quotational and reference accuracy in surgical journals - a continuing peer-review problem.*Journal of the American MedicalAssociation*, 263(10), pp. 1353-4.

Fabbris, L., Boccuzzo, G., Martini, M.C., (a cura di), 2008. *Professionalità nei servizi innovativi per studenti universitari*. Padova, CLEUP.

Fabbris, L., Gnaldi, M., 2008. Indicatori di valutazione della qualità della ricerca negli atenei: sensibilità, sostituibilità e capacità discriminatoria. inFabbris, L., Boccuzzo, G., Martini, M.C., (a cura di), 2008. *Professionalità nei servizi innovativi per studenti universitari*. Padova, CLEUP, pp. 139-171.

Fasanella, A., 1993. *Concettualizzazione e spiegazione sociologica. Il modello nomologico-inferenziale e la sua applicabilità alle scienze sociali*. Milano, Franco Angeli.

Fasanella, A., 2010. Note su realismo e ricerca sociale. *Sociologia e ricerca sociale*, 31(91), pp. 5-42.

Fasanella, A., 2013. Valutazione e validazione: qualche considerazione sulla VQR 2004-2010. *Sociologia e Ricerca Sociale*, 33(100), pp. 132-147.

Fasanella, A., Allegra, S., 1995. Validità dei dati e approccio multritratto-multitecnica. *Sociologia e Ricerca Sociale*, 16(47/48), pp. 231-284.

Fasanella, A., Di Benedetto, A., 2014. Luci ed ombre nella VQR 2004- 2010: un focus sulla scheda di valutazione peer nell'Area 14, *Sociologia e ricerca sociale*, 35(104), pp. 59-84.

Ferriss, A. L. 2004. The quality of life concept in sociology. *The American Sociologist*, 35(3), pp. 37-51.

Frudà, L., 1997. Ricerca valutativa, controllo di qualità e innovazione nella pubblica amministrazione e nella gestione dei servizi pubblici. *Studi di Sociologia*, 35(2), pp. 127-167.

Galtung, J., 1967. *Theory and Method of Social Research.* Oslo, UniversitietsForlaget.

Garfield, E., 1955. Citation indexes for science – New dimension in documentation through association of ideas. *Science*, 122(3159), pp. 108-111.

Garfield, E., 1962. Can citation indexing be automated?.*Essays of an Information Scientist*, 1, pp. 84-90.

Garfield, E., 1971. The mystery of the transposed journal lists – wherein Bradford's low of scattering is generalized according to Garfield's law of concentration. *Current Content*, 3(33), pp. 5-6.

Garfield, E., 1979a. Is citation analysis a legitimate evaluation tool? *Scientometrics* 1(4), pp. 359-375.

Garfield, E., 1979b. *Citation Indexing. Its Theory and Application in Science, Technology, and Humanities*. New York, Wiley.

Garfield, E., 1986. Journal impact vs. influence: A need for 5-year impact factors. *Information Processing & Management*, 22(5), p. 445.

Garfield, E., 1990. How ISI Selects Journals for Coverage: Quantitative and Qualitative Considerations, *Current Contents*, 28(22) pp. 5-13.

Garfield, E., 1994. The Thomson Reuters *Impact factor*. *Current Contents print edition*, 20 June 1994.

Garfield, E.,1996. How can *impact factor*s be improved?.*British Medical Journal*, 313(7054), pp. 411-413.

Garfunkel, J.M., Ulshen, M.H., Hamrick, H.J., Lawson, E.E., 1994. Effect of institutional prestige on reviewers' recommendations and editorial decisions. *JAMA: the Journal of the American Medical Association,* 272(2), pp. 137-138.

Georghiou, L., Howells, J., Rigby, J., Glynn, S., Butler, J., Cameron, H., Cunningham, P., Ausadamongkol, K., Thomas, D., Salazar, A., Barker, K.,Reeve, N., 2000. *Impact of the Research Assessment Exercise and the Future of Quality Assurance in the Light of Changes in the Research Landscape.* Manchester, PREST, University of Manchester.

Geuna, A., Martin, B., 2003. University Research Evaluation and Funding: An International Comparison. *Minerva*, 41(4), pp. 277-304.

Gilbert, J.R., Williams, E.S., Lundberg, G.D., 1994. Is there gender bias in JAMA's peer review process? *JAMA: the Journal of the American Medical Association*, 272(2), pp. 139-142.

Gilbert, N.G., 1977. Referencing as persuasion. *Social studies of science,* 7(1), pp. 112-122.

Glänzel, W., 1988. Characteristics scores and scales in assessing citation impact. *Journal of Information Science*, 14(2), pp. 123-127.

Glänzel, W., 1996. The need for standards in bibliometric research and technology. *Scientometrics*, 35(2), pp. 167-176.

Glänzel, W., 2008. Seven myths in bibliometrics: about facts and fiction in quantitative science studies. inKretschmer, H., Havemann, F., (a cura di), 2008. *Proceedings of WIS 2008, Berlin, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*:http://www.collnet.de/ /Berlin-2008/Glänzel WIS2008smb.pdf.

Glänzel, W., Debackere, K., Thijs, B.,Schubert, A., 2006. A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*, 67(2), pp. 263-277.

Glänzel, W., Moed, H.F., 2013. Opinion paper: thoughts and facts on bibliometric indicators. *Scientometrics*, 96(1), pp. 381-394.

Glänzel, W., Schoepflin, U., 1995. A bibliometric study on ageing and reception processes of scientific literature. *Journal of information Science*, 21(1), pp. 37-53.

Glänzel, W., Schubert, A., 2003. A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), pp.357-367.

Glänzel, W., Thijs, B. 2004. Does co-authorship inflate the share of self-citations? *Scientometrics*, 61(3), pp. 395-404.

Glänzel, W., Thijs, B., Schlemmer, B., 2004. A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, 59(1), pp. 63-77.

Glaser, B.G.,Strauss, A.L., 1967. *The discovery of grounded theory: strategies for qualitative research*. Chicago, Aldine; tr. it. 2009. Strati, A., (a cura di),*La scoperta della GroundedTheory: strategie per la ricerca qualitativa*. Roma, Armando Editore.

Gläser, J., Laudel, G., 2006. Advantages and dangers of 'remote' peer evaluation. *Research Evaluation* 14(3), pp. 186-198.

Gläser, J., Laudel, G., Hinze, S., Butler, L., 2002. *Impact of evaluation-based founding on the production of scientific knowledge: what to worry about and how to find out*, 31, Expertise for the German Ministry for Education and Research.

Godlee, F., Gale, C.R., Martyn, C.N., 1998. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports. *JAMA: the Journal of the American Medical Association*, 280(3), pp. 237-240.

Gómez-Núñez, A.J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., Chinchilla-Rodríguez, Z., 2014. Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, 8(2), pp. 369-383.

Gómez-Núñez, A.J., Vargas-Quesada, B., de Moya-Anegón, F., Glänzel, W., 2011. Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, 89(3), pp. 741-758.

Gonzalez-Pereira, B., Guerrero-Bote, V.P., Moya-Agnegon, F., 2010. A new approach to the metric of journals' scientific prestige: the SJR indictor. *Journal of Informetrics*, 4(3), pp. 379-391.

Goode, W.J., Hatt, P.K., 1952. *Methods in social research*. New York, McGraw Hill; tr. it. 1962.Metodologiadellaricercasociale. Bologna, IlMulino.

Grant, J., Burden, S., Breen, G., 1997. No evidence of sexism in peer review. *Nature*, 390(6659), pp. 438-438.

Guala, C., 2000. *Metodi della ricerca sociale: la storia, le tecniche, gli indicatori*. Roma, Carocci.

Hackett, E. J., Chubin, D. E., 2003. Peer review for the 21st century: applications to education research. in*Peer review of education research grant applications. Implications, considerations, and future directions*, edited by National Research Council. Washington, DC, USA.

Harzing, A.W., 2007. *Publish or Perish*, available from http://www.harzing.com/pop.htm

Harzing, A.W., 2013. Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences?, *Scientometrics,* 93(1), pp. 23-34.

Hempel, C.G., 1952. *Fundamentals of Concept Formation in Empirical Science. in International Encyclopedia of Unified Sciences*, Vol. II. Chicago, University of Chicago Press; tr. it. 1976. *La formazione dei concetti e delle teorie nella scienza empirica*. Milano, Feltrinelli.

Hendrick, C., 1976. Editorial comment. *Personality and Social Psychology Bulletin*, 2(3), pp. 207-208.

Hicks, D., 1999. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), pp. 193-215.

Hicks, D., 2004. The four literatures of social science. inMoed, H., Glänzel, W., Schmoch, U., (a cura di), 2006. *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht, Kluwer Academic, pp. 476-496.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., Rafols, I., 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), pp. 429-431.

Hippler, H.J., Schwarz, N., Sudman, S., (a cura di), 1987. *Social Information Processing and Survey Methodology*. New York, Springer-Verlag.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46), pp.16569-16572.

Hojat, M., Gonnella, J.S., Caelleigh, A.S., 2003. Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. *Advances in Health Sciences Education*, 8(1), pp. 75-96.

Hug, S.E., Ochsner, M., 2014. A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal for Education Law and Policy*, 10(1), pp. 55-68.

Jacsó, P., 2005. As we may search-Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9), pp. 1537-1547.

Jacsó, P., 2009. Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. *Online Information Review*, 33(6), pp. 1189-1200.

Jacsó, P., 2013. The need for end-user customization of the journal-sets of the *subject categories* in the *SCImago Journal Ranking* database for more appropriate league lists. A case study for the Library & Information Science field. *El Profesional de la Informacion*, 22(5), pp. 459-473.

Janssens, F., Glänzel, W., De Moor, B., 2008. A hybrid mapping of information science. *Scientometrics*, 75(3), pp. 607-631.

Jarwal, S.D., Brion, A.M., King, M.L., 2009. Measuring research quality using the journal *impact factor*, citations and 'Ranked Journals': blunt instruments or inspired metrics?.*Journal of Higher Education Policy and Management*, 31(4), pp. 289-300.

Jayasinghe, U.W., Marsh, H.W., Bond, N., 2003. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(3), pp. 279-300.

Jimenez-Contreas, E., de Moja Anegon, F., Lopez-Cozar, E.D., 2003. The evolution of research activity in Spain. The impact of the National Commission for the Evaluation of Research Activity (CNEAI). *Research policy*, 32(1), pp. 123-142.

Justice, A.C., Cho, M.K., Winker, M.A., Berlin, J.A., Rennie, D., 1998. Does masking author identity improve peer review quality?. *JAMA: the Journal of the American Medical Association*, 280(3), pp. 240-242.

Kaplan, A., 1955. Definition and Specification of Meaning. inLazarsfeld, P.F., Rosenberg M., (a cura di), 1955. *The Language of Social Research. A Reader in the Methodology of Social Research*. New York, The Free Press.

Kassirer, J.P., Campion, E.W., 1994. Peer-Review - Crude and Understudied, but Indispensable. *Journal of the American Medical Association*, 272(2), pp. 96-97.

Knorr, K.D., 1978. The nature of scientific consensus and the case of the social sciences. *International Journal of Sociology*, *8*(1/2), pp. 113-145.

Knorr-Cetina, K., 1981. *The manufacture of knowledge. An Essay on the Constructivist and Contextual Nature of Science*. Oxford, Pergamon Press.

Kostoff, R.N., 1995. Federal research impact assessment - axioms, approaches, applications. *Scientometrics*,34(2), pp. 163-206.

Kretschmer, H., Havemann, F., (a cura di), 2008. *Proceedings of WIS 2008, Berlin, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*. http://www.collnet.de/ /Berlin-2008/Glänzel WIS2008smb.pdf.

Krippendorff, K., 1980. *Content Analysis. An Introduction to its Methodology*. London, Sage; tr. it. 1983.Analisi del contenuto. Introduzionemetodologica. Torino, Eri.

Kuhn, T.S., 1962. *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press (1970, 2nd edition, with postscript).

Kuhn, T.S., 1977. *The essential tension. Selected studies in scientific tradition and change*. Chicago, University of Chicago Press.

La Rocca, C., 2013. Commisurare la ricerca. Piccola teleologia della neovalutazione. *Aut Aut*, 360, pp. 69-108.

Labovitz, S., 1970. The assignment of numbers to rank order categories. *American Sociological Review*, 3(3), pp. 515-524.

Lane, D., 2008. Double-blind review: Easy to guess in specialist fields. *Nature*, 452(7183), p. 28.

Langfeldt, L., 2006. The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research evaluation*, 15(1), pp. 31-41.

Larivière, V., Archambault, É.,Gingras, Y., Vignola-Gagné, É., 2006. The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), pp. 997-1004.

Latorur, B., Woolgar, S., 1979. *Laboratory life: The Social Construction of Scientific Facts*. London, Sage.

Laudel, G., 2006. Conclave in the Tower of Babel: how peers review interdisciplinary research proposals. *Research* Evaluation, 15(1), pp. 57-68.

Lazarsfeld, P.F., 1958. Problems in Methodology. in Merton, R.K., Broom, L., Cottrell, L.S., (a cura di), 1958. SociologyToday. New York, Basic Books, pp. 39-78; tr. it. 1967. Problemi di metodologia. inLazaesfeld P.F., 1967. *Metodologia e ricerca sociologica,* (a cura di Capecchi, V.). Bologna, Il Mulino.

Lazarsfeld, P.F., 1967. *Metodologia e ricerca sociologica*,(a cura di Capecchi, V.). Bologna, Il Mulino.

Lazarsfeld, P.F., 1969, Dai concetti agli indici empirici, in Boudon, R., Lazarsfeld, P.F., *L'analisi empirica nelle scienze sociali. Volume I: dai concetti agli indici empirici*. Bologna, Il Mulino.

Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B., 2013. Bias in peerreview. *Journal of the AmericanSociety for Information Science and Technology*, 64(1), pp. 2-17.

Leydesdorff, L., 1998. Theories of citation. *Scientometrics*, 43(1), pp. 5-25.

Leydesdorff, L., 2009. How are new citation-based journal indicators adding to the bibliometric toolbox?.*Journal of the American Society for information Science and Technology*, 60(7), pp. 1327-1336.

Leydesdorff, L., de Moya-Anegón, F., Guerrero-Bote, V.P., 2010. Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for information Science and Technology*, 61(2), pp. 352-369.

Leydesdorff, L., Rafols, I., 2009. A global map of science based on the ISI *subject categories*. *Journal of the American Society for Information Science and Technology*, 60(2), pp. 348-362.

Liefner, I., 2003. Funding, Resources Allocation, and Performance in Higher Education Systems. *Higher Education*, 46(4), pp. 469-489.

Lindsey, D., 1978. *The scientific publication system in social science*. San Francisco, Jossey-Bass.

Lindsey, D., 1988. Assessing precision in the manuscript review process: A little better than a dice roll. *Scientometrics*, 14(1/2), pp. 75-82.

Lloyd, M. E. 1990. Gender factors in reviewer recommendations for manuscript publication. *Journal of appliedbehavior analysis*, *23*(4), pp. 539-543.

Lombardo, C., 1994. *La congiunzione inespressa. I criteri di selezione degli indicatori nella ricerca sociale*. Milano, Franco Angeli.

López-Illescas, C., de Moya-Anegón, F., Moed, H.F., 2008. Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, 2(4), pp. 304-316.

Lotka, A.J., 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences,* 16(12), pp. 317-324.

Lutynski, J., 1988. Un centro di ricerca sulle tecniche di raccolta dei dati. in Marradi, A., (a cura di), 1988. *Costruire il dato*. Milano, Franco Angeli, pp. 117-132.

Macilwain, C., 2010. Wild goose chase. *Nature*, 463(7279), p. 291.

Madge, J., 1962. *The Origins of Scientific Sociology*. New York, The Free Press of Glancoe; tr. it. 1996. *Lo sviluppo dei metodi di ricerca empirica in sociologia*. Bologna, IlMulino.

Mahoney, M.J., 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2), pp. 161-175.

Mansilla, R., Koppen, E., Cocho, G., Miramontes, P., 2007. On the behavior of journal *impact factor* rank order distribution. *Journal of Informetrics*, 1(2), pp. 155-160.

Marradi*,* A., (a cura di), 1988. *Costruire il dato*. Milano, Franco Angeli.

Marradi, A., 1980. *Concetti e metodi per la ricerca sociale*. Firenze, La Giustina.

Marradi, A., 1990a. Classificazioni, tipologie, tassonomie. in AA. VV., 1990, *Enciclopedia delle scienze sociali*. Roma, Istituto dell'Enciclopedia Italiana, pp. 22-30.

Marradi, A., 1990b. Fedeltà di un dato, affidabilità di una definizione operativa. *Rassegna Italiana di Sociologia*, 31(1), pp. 55-96.

Marradi, A., 1991a. Concetti e metodi per la ricerca sociale. in Cardano, M., Miceli, R., (a cura di), 1991. *Il linguaggio delle variabili. Strumenti per la ricerca sociale*. Torino, Rosenberg &Sellier, pp. 17-120.

Marradi, A., 1991b. Misurazione e scale: qualche riflessione e una proposta in Cardano, M., Miceli, R., (a cura di), 1991. *Il linguaggio delle variabili. Strumenti per la ricerca sociale*. Torino, Rosenberg &Sellier, pp. 151-192.

Marradi, A., 2007. *Metodologia della ricerca sociale*. Bologna, Il Mulino.

Marradi, A., Gasperoni, G., (a cura di), 1992. *Costruire il dato 2: vizi e virtù di alcune tecniche di raccolta delle informazioni*. Milano, Franco Angeli.

Marsh, H.W., Ball, S., 1981. Interjudgmental reliability of review for the Journal of Educational Psychology. *Journal of Educational Psychology*, 73(6), pp. 872-880.

Martinotti, G., 1997. *Autonomia didattica e innovazione dei corsi di studio di livello universitario e post-universitario*. Roma, Ministero dell'Università e della Ricerca scientifica.

Mauceri, S., 2003. *Per la qualità del dato nella ricerca sociale. Strategie di progettazione e conduzione dell'intervista con questionario*. Milano, Franco Angeli.

Mayo, E. 1933. *The Human Problems of an Industrial Civilization*. New York,The Macmillan Company; tr. it. 1968. *I problemi umani e socio-politici della civiltà industriale*. Torino, Utet.

McNay, I., 2003. Assessing the assessment: an analysis of the UK Research Assessment Exercise, 2001, and its outcomes, with special reference to research in education. *Science and Public Policy*, 30(1), pp. 47-54.

McNutt R.A., Evans A.T., Fletcher R.H., Fletcher S.W., 1990. The effects of blinding on the quality of peer review. A randomized trial. *JAMA: the Journal of the American Medical Association*, 263(10), pp. 1371-1376.

Meho, L.I., Yang, K. 2007. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13), pp. 2105-2125.

Merton, R.K., 1938. Social structure and anomie. *American Sociological Review*, 3(5), pp. 672-682.

Merton, R.K., 1968*. Social Theory and Social Structure*. Glencoe, Ill; tr. it. 2000. *Teoria e strutturasociale.* Bologna, IlMulino.

Merton, R.K., 1968a. The Matthew effect in science. *Science*, 159(3810), pp. 56-63.

Merton, R.K., 1973. *The sociology of science: Theoretical and Empirical Investigations.* Chicago, University of Chicago Press.

Merton, R.K., 1988. The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis,* 79(4), pp. 606-623.

Merton, R.K., Fiske, M.O., Kendall, P.L., 1956. *The Focused Interview*. New York, The Free Press.

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2), pp. 81-97.

Moed, H., 1996. Differences in the construction of SCI based bibliometric indicator among various producers: a first overview. *Scientometrics*, 32(2), pp. 177-191.

Moed, H., 2002. The *impact factor*s debate: the ISI's uses and limits. *Nature*, 415(6873), pp. 731.732.

Moed, H., 2008. UK Research Assessment Exercises: Informed judgments on research quality or quantity?. *Scientometrics*, 74(1), pp. 153-161.

Moed, H., 2010. Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), pp. 265-277.

Moed, H., Colledge, L., Reedijk, J., Moya-Anegon, F., Guerrero-Bote, V., Plume, A., Amin, M., 2012. Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics*, 92(2), pp. 367-376.

Moed, H., De Bruin, R.E., van Leeuwen, T., 1995. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33(3), pp. 381-422.

Moed, H., Glänzel, W., Schmoch, U., (a cura di), 2006. *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht, Kluwer Academic.

Moed, H., van Leeuwen, T., 1996. *Impact factor*s can mislead. *Nature*, 381(6579), p. 186.

Moed, H., van Leeuwen, T., Reeduk J.A., 1999. Towards appropriate indicators of journal impact. *Scientometrics*, 46(3), pp. 575-589.

Moed, H., Visser, M.S., 2008. Appraisal of citation Data sources. *A report on a study within the Framework Agreement for the development of a new research assessment and founding system commissioned by HEFCE*. www.hefce.ac.uk/pubs/.

Moody, J*.,* 2004. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, 69(2), pp. 213-238.

Neff, B.D., Olden, J.D., 2010. Not so fast: Inflation in *impact factor*s contribute to apparent improvements in journal quality. *BioScience*, 60(6), pp. 455-459.

Nobile, S., 1997. *La credibilità dell'analisi del contenuto*. Milano, Franco Angeli.

Nobile, S., 2008. *La chiusura del cerchio. La costruzione degli indici nella ricerca sociale*. Acireale-Roma, Bonanno.

Norris, M., Oppenheim, C., 2007. Comparing alternatives to the “Web of Science” for coverage of the social sciences’literature. *Journal of Informetrics*, *1*(2), pp. 161-169.

Noruzi, A., 2005. Google Scholar: The new generation of citation indexes. *Libri*, 55(4), pp. 170-180.

Nowak, S., 1976. *Understanding and prediction*. Dordrecht, Reidel.

O'Brien, R. M., 1985. The relationship between ordinal measures and their underlying values: Why all the disagreement?. *Quality & Quantity*, 19(3), pp. 265-277.

Olgden, T.L., Barley, D.L., 2008. The Ups and Downs of Journal *Impact factor*. *Annals of Occupational Hygiene Society*, 52(2), pp. 73-82.

Osuna, C., Cruz Castro, L., SanzMenédez, L., 2010. *Knocking down some Assumptions about the Effects of Evaluation Systems on Publications*, Instituto de Políticas y BienesPúblicos (IPP), CHS-CSIC, Working Paper, Number 10.

OVSU, 1997a. *Relazione sull'attività svolta nel 1996*, Doc 7/97.

OVSU, 1997b. *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.

OVSU, 1998. *Indicazioni per la preparazione delle relazioni dei Nuclei di valutazione interna e griglia minima di indicatori*, Doc 11/98.

OVSU, 1999a. *Programma di valutazione istituzionale delle università Programma VIU*, Doc 10/99, aprile 1999.

OVSU, 1999b. *Programma di valutazione della produzione scientifica delle università Programma VPS*, Doc 10/99, febbraio 1999.

Owens, B., 2013. Long-term research: Slow science. *Nature*, 495(7441), pp. 300-303.

Oxman A.D., Guyatt, G.H., Singer, J., 1991. Agreement among reviewers of review articles. *Journal of ClinicalEpidemiology*, 44(1), pp. 91-98.

Palumbo, M., 2001. *Il processo di valutazione. Decidere, programmare, valutare*. Milano, Franco Angeli.

Palumbo, M., 2013. Chi ha paura della valutazione cattiva?.*Sociologia e Ricerca Sociale*, 33(100), pp. 52-65.

Pasquinelli, A., 1977. *Nuovi principi di epistemologia*. Milano, Feltrinelli.

Pathak, L.P., 2000. *Sociological Terminology and Classification Schemes*. New Delhi, Mittal Publications.

Peters, D.P., Ceci, S. J., 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(02), pp. 187-195.

Pitrone, M.C., 1995. La formulazione delle domande, alcuni problemi metodologici. *Sociologia e ricerca sociale*, 56(16), pp. 46-76.

Pitrone, M.C., 2009. Sondaggi e interviste. *Lo studio dell'opinione pubblica nella ricerca sociale.* Milano, Franco Angeli.

Polanyi, M., 1962. *The tacit dimension*. London, Routledge & Kegan Paul.

Preston, C.C., Colman, A.M., 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Actapsychologica*, 104(1), pp. 1-15.

Pudovkin, A.I., Garfield, E., 2002. Algorithmic procedure for finding semantically related journals, *Journal of the American Society for Information Science and Technology*, 53(13), pp. 1113-1119.

Radicchi, F., Fortunato, S., Castellano, C., 2008. Universality of citation distributions: toward an objective measure of scientific impact. *Proceedings of the national Academy of Sciences*, 105(45), pp. 17268-17262.

Rafols, I., Leydesdorff, L., 2009. Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, 60(9), pp. 1823-1835.

Rapley, M., 2003. *Quality of life research: A critical introduction*. London, Sage.
Reale, E., (a cura di), 2008. *La valutazione della ricerca pubblica: una analisi della valutazione triennale della ricerca*. Milano, Franco Angeli.
Reale, E., 2013. La valutazione della ricerca e il cambiamento nelle università. *Sociologia e Ricerca Sociale*, 33(100), pp. 148-159.
Reale, E., Barbara, A., Costantini, A., 2007. Peer review for the evaluation of academic research: lessons from the Italian experience. *Research Evaluation*, 16(3), pp. 216-228.
Rebora, G., 2012. Venti anni dopo. Il percorso della valutazione dell'università in Italia e alcune proposte per il futuro. *Liuc Papers* n. 257*,* SerieEconomiaaziendale 38*.*
REF 01.2012. *Assessment framework and guidance on submissions*. http://www.ref.ac.uk/pubs/2011-02/.
REF 01.2012. *Panel criteria and working methods*. http://www.ref.ac.uk/pubs/2012-01/.
Ricolfi, L., 1992. Sul rapporto di indicazione: l'interpretazione semantica e l'interpretazione sintattica. *Sociologia e Ricerca Sociale*, 13(39), pp. 57-79.
Rizzi, D., Silvestri, P., 2002. *La valutazione del sistema universitario italiano: una storia recente*. Modena, CAPP Centro di Analisi delle Politiche Pubbliche.
Robinson, W., 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), pp. 351-357.
Rockwell, S., 2005. *Ethics of* peer review*: a guide for manuscript reviewers*. New Have, Office of research integrity, US. Department of Health and Human Services.
Rodrìguez-Ruiz, O., 2009. The citation indexes and the quantification of knowledge. *Journal of Educational Administration*, 47(2), pp. 250-266.
Rosenthal, R.A., Jacobson, L., 1968. *Pygmalion in the classroom. Teacher expectation and pupils'intellectual development*. New York, Rinehart and Winston; tr. it. 1992.*Pigmalione in classe: aspettative degli insegnanti e sviluppo intellettuale degli allievi*. Milano, Franco Angeli.
Rousseau, R., 2009. On the relation between the WoS*impact factor*, the Eigenfator, the SCImago Journal Rank, the Article Influence Score and the journal h-index. http://eprints.rclis.org/16448/ 30.06.2014.
Sandström, U., Hällsten, M., 2008. Persistent nepotism in peer-review. *Scientometrics*, 74(2), pp. 175-189.
Sartori, G., (a cura di), 1984. *Social Science Concepts. A Systematic Analysis*. Beverly Hills, London, New Dehli, Sage.
Sartori, G., 1984. Guidelines forConcept Analysis. inSartori, G., (a cura di), *Social Science Concepts. A Systematic Analysis*. Beverly Hills, London, New Dehli, Sage, pp. 15-85.
Scarr, S., Weber, B.L.R., 1978. The reliability of reviewers for the American Psychologist. *American Psychologist*, 33(10), pp. 935.
Schneider, J.W., 2009. An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science*, 8(3), pp. 364-378.
Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., Smith, R., 2004. Effects of training on quality of peer review: randomised controlled trial. *BMJ: British Medical Journal*, 328(7441), p. 673.
Schutz, W.C., 1958. On categorizing qualitative data in content analysis. *Public Opinion Quarterly*, 22(4), pp. 503-515.
Schwarz, N., Hippler, H.J., 1987. What Response Scales May Tell Your Respondents. in Hippler, H.J., Schwarz, N., Sudman, S., (a cura di), 1987. *Social Information Processing and Survey Methodology*. New York, Springer-Verlag, pp. 163-178.
Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., Clark, L., 1991. Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55(4), pp. 570-582.
Scopus, 2012. *Content Coverage Guide*. http://www.elsevier.com/__data/assets/pdf_file/0019/148402/contentcoverageguide-jan-2013.pdf

Scopus, 2014. Helping to improve the submission & success process for Editors & Publishers. *Scopus Journal FAQs*, http://www.elsevier.com/__data/assets/pdf_file/0017/234332/SC_FAQ-content-selection-process-22092014.pdf

Scott, A., 2007. Peer review and the relevance of science. *Futures*, 39(7), pp. 827-845.

Scott, W.A., 1974. Inter*referee* agreement on some characteristics of manuscripts submitted to the Journal of Personality and Social Psychology. *American Psychologist*, 29(9), pp. 698-702.

Seglen, P.O. 1997b. Why the *impact factor*should not be used for evaluating research. *British Medical Journal*, 314(7079), pp. 497-502.

Seglen, P.O., 1997a. Citations and journal *impact factor*s: questionable indicators of research quality. *Allergy*, 52(11), pp. 1050-1056.

Shatz, D., 2004. *Peer review: a critical inquiry*. Lanham, Rowman& Littlefield.

Shubert, A., Braun, T., 1996. Cross-field normalization for scientometric indicators. *Scientometrics*, 36(3), pp. 311-324.

Sirgy, M.J., Michalos, A.C., Ferriss, A.L., Easterlin, R.A., Patrick, D., Pavot, W., 2006. The Qualityity-of-Life (QOL) Research Movement: Past, Present, and Future. *Social Indicators Research*, 76(3), pp. 343-466.

Smith, R., 2006. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4), pp. 178-182.

Starbuck, W.H., 2005. How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, 16(2), pp. 180-200.

Statera, G., 1997. *Logica dell'indagine scientifico-sociale*. Roma, Seam.

Talib, A., 2001. The Continuing Behavioural Modification of Academics since the 1992 Research Assessment Exercise. *Highter Education Review*, 33(3), pp. 30-46.

Thijs, B., Glänzel, W., 2006. The influence of author self-citations on bibliometricmeso-indicators. The case of European universities. *Scientometrics*, 66(1), pp. 71-80.

Thomson Reuters, 2014. Connecting the dots across the research ecosystem. *Thomson Reuters White Papers*, http://wokinfo.com/media/pdf/connecting_the_dots.pdf

Thurstone, L.L., 1928. Attitudes can be measured. *American Journal of Sociology*, 33(4), pp. 529-554.

Tufte, E.R., 1970. *Quantitative analysis of social problems*. Reading, Addison-Wesley.

van Leeuwen, T., Moed, H., 2005. Characteristics of Journal *Impact factor*s: the effects of uncitedness and citation distribution on the understanding of journal *impact factor*s. *Scientometrics*, 63(2), pp. 357-371.

van Leeuwen, T., van der Wurff, L.J., de Craen, A.J.M., 2007. Classification of 'research letters' in general medical journals and its consequences in bibliometric research evaluation processes. *Research Evaluation*, 16(1), pp. 59-63.

Vanclay, J.K., 2009. Bias in the journal *impact factor*. *Scientometrics*, 78(1), pp. 3-12.

van Leeuwen, T., Costas, R., Calero-Medina, C., Visser, M., 2013. The role of editorial material in bibliometric research performance assessments. *Scientometrics*, 95(2), pp. 817-828.

van Leeuwen, T., Moed, H., Tijssen, R.J., Visser, M.S.,van Raan, A.F., 2001. Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, 51(1), pp. 335-346.

van Raan, A., 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36(3), pp. 397-420.

van Raan, A., 2005. Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics,* 62(1), pp. 133-143.

van Raan, A., 2008a. Scaling rules in the science system: influence of field-specific citation characteristics on the impact of research groups. *Journal of the American Society for Information Science and Technology,* 59(4), pp. 565-576.

van Raan, A., 2008b. Self-citations as an impact-reinforcing mechanism in the science system. Journal of the *American Society for Information Science and Technology,* 59 (10) pp. 1631-1643.

van Rooyen, S., Black, N., Godlee, F., 1999. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology*, 52(7), pp. 625-629.

van Rooyen, S., Godlee, F., Evans, S., Black, N., Smith, R., 1999. Effect of open peer review on quality of reviews and on reviewers'recommendations: a randomised trial. *BMJ: British Medical Journal*, 318(7175), pp. 23-27.

van Rooyen, S., Godlee, F., Evans, S., Smith, R., Black, N., 1998. Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA: the Journal of the American Medical Association*, 280(3), pp. 234-237.

Veiera, E.S., Gomes, J.A., 2009. A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, 81(2), pp. 587-600.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., van Leeuwen, T., van Raan, A., Visser, M.S., Wouters, P., 2012. The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), pp. 2419-2432.

Waltman, L., Schreiber, M., 2012. On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64(2), pp. 372-379.

Waltman, L., van Eck, N., 2009. Some comments on Egghe's derivation of the *impact factor* distribution. *Journal of Informetrics*, 3(4), pp. 363-366.

Waltman, L., van Eck, N., 2012. A new methodology for constructing a publication-level classification system of science. Journal *of the American Society for Information Science and Technology*, 63(12), pp. 2378-2392.

Waltman, L., van Eck, N., van Raan, A., 2012. Universality of citation distribution revisited. *Journal of the American Society for Information Science and Technology*, 63(1), pp. 72-77.

Waltman, L., van Eck, N., van Leeuwen, T.,Visser, M.S., 2013. Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, 7(2), pp. 272-285.

Weingart, P., 2005. Impact of bibliometrics upon the science system: inadvertent consequences?. *Scientometrics*, 62(1), pp. 117-131.

Wennerås, C., Wold, A., 1997. Nepotism and sexism in peer-review. *Nature*, 387(6631), pp. 341-343.

Westerheijden, D., 1997. A Solid Base for Decisions: Use of the VSNU research Evaluation in Dutch Universities. *Higher Education*, 33(4), pp. 397-413.

Whitley, R., 2007. The changing governance of the public sciences: the consequences of research evaluation systems for Knowledge production in different countries and scientific fields. in Whitley, R., Glaser, J., (a cura di), 2007. *The changing governance of the sciences. The advent of the Research evaluation systems*. Dordrecht, Springer, pp. 3-27.

Whitley, R., Glaser, J., (a cura di), 2007. *The changing governance of the sciences. The advent of the Research evaluation systems*. Dordrecht, Springer.

Wilson, A.E., 2007. Journal *impact factor* are inflated. *BioScience*, 57(7), pp. 550-551.

Yankauer, A. 1991. How blind is blind review?.*American Journal of Public Health*, 81(7), pp. 843-845.

Zhang, L., Liu, X., Janssens, F., Liang, L., Glänzel, W., 2010. Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), pp. 185-193.

Zipf, G.K., 1949. *Human Behaviour and the Principle of Last Effort*. Cambridge, Wesley.

Zuckerman, H., Merton, R.K., 1973. Patterns of Evaluation of science: Institutionalization, Structure and Functions of the *referee* system. In Merton, R.K., 1973. *The sociology of science: Theoretical and Empirical Investigations.* Chicago, University of Chicago Press, pp. 569-609.