EDITORIAL

Salvatore Lorusso and Vincenzo Barone

The evaluation of scientific research: the result of merit-based or discretionary criteria?

The evaluation of research has assumed a central role in Italy. This is, on the one hand, due to the need to assign distinct parameters as a benchmark for researchers' careers and on the other, the need to identify an efficient systemic indicator for allocating financial resources. Among the methods used, bibliometric tools occupy a prominent place. These are the product of mathematical and statistical techniques useful in analysing the distribution patterns of scientific publications and exploring their impact within the scientific community.

In principle, this kind of qualitative tool can contribute to create an appropriate and fair system of evaluation which can be used in a large number of specific contexts, from public contributions to competitions, and also in the classification of single research products (journals, articles, books, monographs).

Moreover, the definition and sharing of evaluation mechanisms that depend (in part) on numerical indices can assume a strongly symbolic significance: it demonstrates that an entire research community is willing to make an effort to substantiate its own credibility, adopting in this process criteria that can be presented and explained in detail even to observers that do not belong to the same community.

In order to analyse the situation in more detail, one can focus on two different aspects: the aims of an evaluation process, and the actual procedures used in achieving these aims. Based on this distinction, the relationship between merit-based and discretionary criteria can then be explored.

A system of meritocracy is one in which motivation, talent and commitment are recognised as individual values and qualities which, under conditions of equal opportunity, produce benefits for the whole community. To function well however, a merit system requires "metrics", in other words, clear, evident markers for assessing and rewarding merit.

Meritocratic culture in Italy has fragile roots. In the seventies the education sector was seriously affected by a situation which did not acknowledge any form of meritocracy, leading to a delegitimization of the very concept of "meritocratic selection". The setup and adoption of evaluation mechanisms capable of identifying and rewarding merit represents at present the only antidote against cronyism and favouritism bestowed on family members.

Given such an aim, the question arises as to how one should go about achieving it. In other words, how much discretionary power should be accepted/tolerated in an evaluation process that correctly aims at recognizing merit? And conversely, how much weight should be given to more quantitative numerical indices?

Once the question is formulated in this manner, it becomes clear that the answer must necessarily depend on the specific context. Thus, for example, the degree of

physiological discretion will probably be greater for historical, humanist and economic sciences, but lower for technical and experimental sciences.

Moreover, procedures aiming at identifying merit must be adapted to the specific type of evaluation: selecting a Ph.D. candidate is an entirely different issue to choosing a senior scientist to direct a research laboratory; in the selection of projects for funding, different considerations will enter into play in strongly applicative areas, as opposed to methodological developments. Finally, none of the available bibliometric tools is reliable enough to be used in isolation. As a consequence, an evaluation procedure that combines transparency and verifiability, with adequate flexibility and wide applicability, will have to rely on a blend of complementary elements, taken in different proportions depending on the specific context.

What are, then, the single evaluation tools at our disposal?

The most well-known bibliometric tool is undoubtedly the *Impact Factor* (IF), owned by Thomson Reuters (ISI - Institute for Scientific Information), which collects data from over 14,000 journals present in its web portal: it is an evaluation system which determines the frequency with which an article is cited in a given period.

Other widely adopted tools include:

- *Peer Review*, a quality indicator typically used in selecting articles for publication;
- *Open Linking*, a technique with a secure future in the field of electronics, is a reference service offered by aggregators, which transforms citations into hyperlinks, allowing researchers to navigate online from one article to another regardless of the journal and publishing house;
- *H Factor or H index* which indicates the real scientific contribution of a researcher.

Peer Review

An interesting perspective can be gained by examining the procedure that is typically followed in the evaluation of a specific class of scientific products, namely, articles submitted for publication in journals – what is the methodological path that leads to deciding what to accept or what not to accept?

For prestigious journals – *Nature and Science*, for example – the first decision is taken by the scientific committee. After this initial screening, articles are sent to experts in the field, “referees”: a peer review – those who judge today, will be judged tomorrow. Although rather than equals, it is a review by people competing against each other, protected by strict anonymity, and as such, represents a much criticised system. It should be noted that experts need to instruct “referees” in suggesting amendments to improve the product, so that it will subsequently be published. Certain behaviour, not always considered proper, can at times be dictated by personal motives related to the past when the evaluator himself was the object of assessment: objectively, such situations make it more difficult to comply with the principles of meritocracy and scientific fairness.

The Impact Factor

As already noted, the impact factor is an index that measures the frequency with which articles in a journal are cited in a particular year or period. It is calculated by counting the total number of citations received in the current year of all the articles published in that journal during the previous two years, and dividing the result by the total number of articles published in the journal during these two years. The index takes into account only citations addressed to journals that have already obtained IF rankings.

Despite its prevalence, the IF is a somewhat questionable tool. Its detractors criticise it because many authors cite for convenience and not for any credit assigned to the article. Others develop citation networks to benefit privileged members with connections to scientific schools. Lastly, there is self citation, which represents the most obvious limitation to the IF. This system is nevertheless one of the few tools that allows research to be evaluated on a more solid basis than just individual assessment. As a result its use has invaded other areas outside academia: publishers and information providers use it as a privileged instrument of market research, while universities and ministries calculate the productivity and prestige of researchers, identifying the quality of structures and arranging their academic and institutional policies accordingly. There are those who claim that the IF is a useful tool but not the most reliable. Andrea Lenzi, president of the "*Consiglio Universitario Nazionale*" or CUN (National University Council), bases this assumption on the fact that: "Italian universities are willing to be evaluated, as demonstrated by the launching of the national registry of professors in which their activities during the year are made public. It is also true however that the IF is only one of the many rating systems used today. Naturally, there are many who criticize it because it has more than a few limitations:

- the only citations considered are those found in several journals; citations in books, conference reports, theses, etc. are not included;
- the analysis aims at counting citations, without entering into the merit of their contents: as an extreme case, but not improbable, one could be cited numerous times for scientifically incorrect works".

The founder of the IF himself, Eugene Garfield, advised against using this index, because the evaluation of each individual scientific work, which should always represent the real end purpose, depends on the reputation and circulation of the journal in which it is published. If in fact a first class paper appears in a lesser known or less widespread journal, it is at a disadvantage, compared to a mediocre paper which appears in a journal of greater prestige.

In short, it is advisable to consider registering a journal with ISI, so that its data is made known and can be examined; the interpretation of such data should not, however, be considered unduly binding.

Roberto Cafferata, president of the *Accademia Italiana di Economia Aziendale* – AIDEA (Italian Academy of Business Economics), also raises some doubts about the efficiency of the IF: "This system has been used in the US and other countries for some time to determine the value of scientific journals and to quantify the quality of individual publications. Over the years however, the IF has become a subject of controversy both in medicine and economics. In our sector, the most suitable system is considered to be Peer Review used in conjunction with other bibliometric indices. Peer Review, essentially allows there to be an international standard evaluation; it is a starting point,

but then other systems for assessment may be introduced to analyse the merits of product content".

As hinted before, Peer Review is the standard criteria used to guarantee the quality of what journals publish – and this applies not only to those that have a director, who acts as referee in the choice of papers, but also for those with large numbered steering committees and editorial staff, consisting of scholars and experts, some of whom have a prestigious reputation.

One wonders: "Does the reading of papers, requested in a discretionary and private way of guarantors and third parties, effectively ensure against the influence of schools, interest groups or other forms of pressure, whether cultural, academic or otherwise? And why should a well-written, thorough and precise work published in journals without any particular requirements be regarded as being inferior compared to another paper published in accredited journals according to certain requirements?

The h-index factor

As already suggested, the IF is not the only existing bibliometric index. Generally used as an indicator of scientific productivity, the h-index takes its name from its creator, Jorge E. Hirsch of the University of California at San Diego. The h-index is of great significance, since it provides an estimate of the influence of a scientist on the community regardless of individual papers of great success or works by authors who, despite having published a great deal, have produced articles of little interest, as happens with the IF. A high h-index can represent a factor of prestige within the scientific community.

The observations and the different viewpoints that have been reported lead to several considerations. The first comes from an aspect prevalent today, known as the "colonisation of communication", in other words the exceedingly widespread need to know about details and conditions related to different activities and respective operators: an aspect which includes the scientific community and its followers.

The second stems from the "quantification of quality", in other words metrology. The adoption of metrology with the goal of evaluating scientific dignity gives rise to many controversial issues, and leads to a perception of one-sidedness and bureaucracy that is not conducive to achieving broader objectives.

Nevertheless, and especially in the sense of recovering the cultural credibility mentioned at the beginning, introducing a significant component of evaluation based on quantitative criteria (mainly bibliometric) must now be considered indispensable. To refuse to use it would inevitably mean endorsing arbitrary evaluation practices, which are totally unacceptable, but sadly all too common.

Reservations which have emerged regarding single bibliometric criteria should therefore be interpreted constructively. In other words, they must be regarded as an incentive for research communities to develop evaluation mechanisms suited to their own specific field by reflecting thoroughly on each individual instance. This obviously means avoiding unjustifiable automatic reactions; adjusting the value of various criteria according to the specific purpose of the evaluation and trying to reconcile numerical data with other indications of quality, possibly similar to those underlying the mechanisms of "peer review".

Once a scientific community has proven willing to be judged in terms of its ethical and scientific reliability, by adopting evaluation criteria that are publicly accessible and easily verifiable, then it will become possible and advisable to shift the accent towards the role of open scientific and cultural debate in monitoring scientific validity.

Put in these terms, this perspective is not only limited to technical, experimental and naturalistic disciplines, but can encompass historical, humanistic and economic ones. This is especially relevant, if one accepts a univocal, rather than a dichotomous vision of culture, particularly evident, though in a more limited way nowadays, in "art and science", namely in studies in the field of cultural and environmental heritage.

La valutazione della ricerca scientifica: risultato di criteri meritocratici o discrezionali?

La valutazione della ricerca in Italia ha assunto un carattere di centralità dovuto, da un lato alla all'intenzione di utilizzare parametri di riferimento specifici per inquadrare la carriera dei ricercatori, dall'altro alla necessità di individuare un indicatore sistemico di efficienza per l'allocazione delle risorse economiche. Fra i metodi utilizzati, un posto di primo piano è occupato dagli strumenti bibliometrici che sono il prodotto di tecniche matematiche e statistiche utili per analizzare i modelli di distribuzione delle pubblicazioni scientifiche e per esplorare il loro impatto all'interno delle comunità scientifiche.

Indicatori numerici di questo tipo possono in linea di principio concorrere alla formulazione di criteri di valutazione efficaci e corretti in innumerevoli contesti pratici, che spaziano dai concorsi pubblici per il reclutamento o per le progressioni di carriera, alla selezione di progetti di ricerca in bandi di finanziamento, fino alla classificazione di singoli prodotti della ricerca (riviste, articoli, volumi, monografie).

L'elaborazione e la condivisione di meccanismi valutativi che contemplino (anche) il ricorso a indicatori numerici riveste inoltre un significato fortemente simbolico: testimonia la disponibilità e lo sforzo, da parte di una comunità disciplinare, di dimostrare e sostanziare la propria credibilità, utilizzando criteri non autoreferenziali, ma chiaramente illustrabili e verificabili anche da parte di coloro che di questa comunità non fanno parte.

Per articolare in maniera più chiara il discorso, è utile identificare e distinguere due aspetti: la finalità della valutazione e il meccanismo adottato per raggiungere questa finalità. Sulla base di questa distinzione, ci si potrà poi interrogare sulla dicotomia meritocrazia /discrezionalità.

La meritocrazia è un sistema in cui motivazione, talento e impegno individuale sono riconosciuti come valori, come doti che producono, in condizioni di eguaglianza di opportunità, vantaggi per tutta la collettività. Per funzionare bene, un sistema meritocratico necessita però di una "metrica", di segnali chiari e visibili che accertino e premino il merito.

In Italia la cultura meritocratica ha radici fragili. Negli anni Settanta il settore dell'istruzione fu investito da un'ondata livellatrice che ha delegittimato il concetto stesso di "selezione meritocratica". Un meccanismo di valutazione capace di identificare e quantificare il merito rappresenta oggi l'unico antidoto contro "clientelismo e parentismo".

Accettata questa finalità, resta però un aspetto "implementativo": in altre parole, in una valutazione che si proponga di identificare e premiare il merito, quanta discrezionalità è auspicabile/tollerabile? Quanto peso devono invece assumere degli indicatori numerici più quantitativi?

Una volta posta la questione in questi termini, la risposta non può che dipendere dal contesto specifico. Così, ad esempio, il margine di fisiologica discrezionalità insita nelle valutazioni è probabilmente maggiore nelle scienze storico-umanistiche ed economiche, minore nell'ambito delle scienze tecniche e sperimentali.

D'altro canto, l'identificazione del merito richiede, all'atto pratico, criteri e procedure notevolmente diversi a seconda della specifica valutazione: altro è selezionare il miglior candidato a un dottorato di ricerca, altro scegliere uno scienziato già esperto cui affidare un intero laboratorio. Analogamente, nel decidere un

finanziamento alla ricerca, gli elementi da prendere in considerazione in un contesto fortemente applicativo non sono gli stessi che se invece si punta a uno sviluppo metodologico innovativo. Ancora, fra i vari indicatori bibliometrici disponibili, nessuno è sufficientemente solido da poter essere utilizzato da solo. Insomma, un approccio valutativo trasparente e verificabile, ma al contempo sufficientemente flessibile e di ampia applicabilità, deve necessariamente far riferimento a una combinazione di più elementi, pesati diversamente a seconda dello specifico contesto.

Quali sono, dunque, i singoli strumenti di valutazione a disposizione?

Lo strumento bibliometrico più noto è senza dubbio l'Impact Factor (IF) di proprietà della Thompson Reuters - Institute for Scientific Information (ISI), che raccoglie i dati di oltre 14 mila riviste presenti nel suo portale web: si tratta di un sistema di valutazione che determina la frequenza con cui un articolo viene citato in un determinato periodo.

Altri strumenti sono:

- *il Peer Review, indicatore di qualità tipicamente utilizzato nella selezione degli articoli pubblicabili;*
- *in ambiente elettronico una tecnica di sicuro avvenire, l'Open Linking, servizio di referenza offerto dagli aggregatori, che trasforma le citazioni in hyperlink e consente ai ricercatori di navigare online da articolo a articolo, indipendentemente dalla rivista e dalla casa editrice;*
- *il Fattore H o h index che punta a quantificare il contributo scientifico complessivo di un ricercatore.*

Peer Review

Indicazioni interessanti si possono ricavare esaminando la prassi comunemente accettata per la valutazione di prodotti scientifici – in particolare, gli articoli da pubblicare su riviste: qual è il percorso metodologico che conduce a decidere cosa accettare e cosa no?

Per prestigiose riviste – Nature e Science, per esempio – la prima decisione è presa dal comitato scientifico. Dopo questo primo vaglio, i lavori vengono inviati a esperti del settore, "referee". Si tratta di una revisione fra pari: chi oggi giudica, domani è giudicato. Anche se, più che fra pari, la revisione avviene in realtà fra persone in competizione fra loro, protette da un rigoroso anonimato; un aspetto che ha dato luogo a numerose critiche. Normalmente, i referee vengono invitati a dare suggerimenti per migliorare il prodotto, di modo che esso possa poi essere pubblicato. Comportamenti non del tutto corretti sono talvolta dettati da motivazioni personali, ad esempio riconducibili a episodi pregressi in cui il referee è stato a sua volta oggetto di valutazione: situazioni che oggettivamente rendono più difficile il rispetto dei principi di meritocrazia e di onestà scientifica.

Impact Factor (IF)

Come già sottolineato, l'Impact Factor è un indice che misura la frequenza con la quale gli articoli di una rivista vengono citati in un particolare anno o periodo. Per il calcolo, si contano le citazioni ricevute nell'anno di riferimento da tutti gli articoli pubblicati nella rivista nei due precedenti anni, e si divide il risultato per il numero totale di articoli comparsi sulla rivista in questi due anni. L'indice tiene conto esclusivamente delle citazioni rivolte a riviste che abbiano già ottenuto valutazioni di IF.

Malgrado la sua diffusione, l'IF è uno strumento piuttosto contestato. I suoi detrattori lo criticano poiché molti autori citano per convenienza e non per il credito assegnato ad un lavoro; altri sviluppano vere e proprie reti citazionali, in modo da realizzare un indebito vantaggio per gli aderenti a specifiche scuole scientifiche; infine, l'autocitazione rappresenta il limite più palese dell'IF. Malgrado ciò, questo sistema è uno dei pochi strumenti che permetta di valutare l'impatto della ricerca su una base più solida della valutazione individuale; per questo motivo, il suo uso è dilagato anche fuori del campo accademico: gli editori e i fornitori di informazioni ne fanno uno strumento privilegiato di ricerca di mercato, mentre le università e i ministeri calcolano la produttività e il prestigio dei ricercatori, giudicano la qualità delle strutture, e orientano di conseguenza le politiche accademiche e istituzionali.

Alcuni esperti considerano l'IF uno strumento utile, ma non il più affidabile. Così, Andrea Lenzi, presidente del Consiglio Universitario Nazionale (CUN), parte da una premessa: «Le università italiane sono disponibili a sottoporsi a valutazione, come dimostra l'avvio dell'anagrafe nazionale dei professori che rende pubbliche le attività svolte durante l'anno. Ma è pur vero che l'IF è uno dei tanti sistemi di valutazione che oggi vengono utilizzati. Certo sono in molti a criticarlo perché presenta non pochi limiti:

- *le citazioni considerate sono solo quelle presenti in alcune riviste; non sono calcolate le citazioni presenti nei libri, negli atti di congressi, nelle tesi, ecc. ;*
- *l'analisi si limita a contare le citazioni,ma non entra nel merito del loro contenuto: come caso limite, ma non inverosimile, si potrebbe essere citati molte volte come autore di lavori scientificamente non corretti».*

Lo stesso inventore dell'IF, Eugene Garfield, ha sconsigliato di utilizzare questo indice, perché esso subordina la valutazione del singolo lavoro scientifico, che dovrebbe rappresentare sempre il fine autentico e ultimo, alla rinomanza e diffusione della rivista in cui esso è pubblicato. Invece un articolo di alto valore,ma pubblicato in una rivista poco nota e scarsamente diffusa, risulta penalizzato rispetto a un articolo mediocre apparso su una rivista di maggiore prestigio.

Insomma, è senz'altro opportuno effettuare la registrazione di una rivista presso l'ISI, di modo che i dati relativi siano resi noti e possano essere esaminati; l'interpretazione di questi dati, tuttavia, non deve avvenire in maniera indebitamente vincolante.

Sull'efficienza dell'IF solleva qualche perplessità anche Roberto Cafferata, presidente dell'Accademia Italiana di Economia Aziendale (AIDEA): «Questo sistema è usato da tempo negli USA e in altri Paesi per stabilire il valore delle riviste scientifiche e per quantificare la qualità delle singole pubblicazioni. Da tempo, però, l'IF è diventato oggetto di controversia per chi si occupa di medicina o di scienze economiche. Per il nostro settore, ad esempio, riteniamo più idoneo il sistema Peer Review accompagnato da altri indici bibliometrici. In sostanza il Peer Review permette di avere una valutazione standard internazionale; è il punto di partenza, ma poi si possono utilizzare altri sistemi di valutazione per analizzare nel merito il contenuto di un prodotto».

Il criterio, come si è già accennato, è ampliamente utilizzato per garantire la qualità di quanto le riviste pubblicano, non solo nei casi in cui un unico direttore svolge di fatto il ruolo di referee nella scelta degli articoli, ma anche nel caso di riviste che hanno folti comitati di direzione e di redazione, costituiti da studiosi e specialisti anche prestigiosi.

E tuttavia ci s può domandare: "La lettura, chiesta in via discrezionale e privata di garanti e terzi, assicura effettivamente contro l'influenza di scuole, interessi, gruppi o forme di pressione culturale o accademica o di altro tipo? E perché un lavoro accurato e corretto, pubblicato su riviste prive di certi requisiti, deve essere considerato meno di un altro di minore valore, apparso su riviste più accreditate in base a tali requisiti?"

Indice H o h index

Come già sottolineato, l'IF non è l'unico indice bibliometrico esistente. Quello più utilizzato come indicatore della produttività scientifica è l'Indice H o" h index", che prende il nome dal suo ideatore Jorge E. Hirsch della University of California di San Diego. L'h index assume grande rilevanza poiché fornisce una stima dell'influenza di uno scienziato sulla comunità prescindendo da singoli articoli di grande successo, o anche dai lavori di autori che, pur avendo pubblicato molto, hanno prodotto solo articoli di scarso interesse, come avviene invece con l'IF. Un h index elevato può rappresentare un fattore di prestigio all'interno della comunità scientifica.

Dalle osservazioni presentate, e dai punti di vista riportati, emergono varie considerazioni. Una prima considerazione riguarda un aspetto oggi imperante, definito "colonizzazione della comunicazione", ovvero la spinta esagerata ed eccessiva a conoscere dati e situazioni inerenti alle diverse attività e ai rispettivi operatori: aspetto che coinvolge anche il mondo scientifico e i suoi adepti.

La seconda è da ricondursi alla "quantificazione della qualità" ovvero alla metrologia. Quando viene utilizzata nell'intento di valutare la dignità scientifica, la metrologia presenta innegabilmente dei lati discutibili, e determina una percezione di unilateralità e burocratizzazione che non giova alla realizzazione di obiettivi più ampi.

Ciononostante, e specialmente nel senso del recupero di credibilità culturale cui accennavamo all'inizio, l'introduzione nei meccanismi di valutazione di una significativa componente basata su criteri quantitativi (essenzialmente bibliometrici) va considerata oramai irrinunciabile. Rifiutarne l'uso in maniera pregiudiziale e aprioristica implicherebbe inevitabilmente un avallo verso pratiche valutative inammissibili e arbitrarie, purtroppo non prive di diffusione.

Le riserve che sono emerse relativamente ai singoli criteri bibliometrici vanno insomma interpretate costruttivamente, come stimolo per le comunità disciplinari a elaborare, attraverso una riflessione rigorosamente svincolata dalle singole istanze valutative, i meccanismi di valutazione più adeguati alle proprie specificità; evitando, come è ovvio, automatismi ingiustificabili; adeguando i pesi dei vari criteri alla finalità specifica della valutazione; e contemperando i dati numerici con altre indicazioni di qualità, eventualmente non dissimili da quelle sottese ai meccanismi di "peer review".

Una volta che una comunità disciplinare abbia dimostrato la propria disponibilità a farsi giudicare, esplicitando i propri meccanismi di valutazione in maniera pubblica e trasparente, e sostanziando in maniera inequivocabile la propria credibilità etica e scientifica, diventa finalmente sostenibile rivendicare il ruolo del libero dibattito scientifico e culturale nel formare e attribuire i giudizi di validità scientifica.

Posta in questi termini, la prospettiva non è limitata alle discipline di carattere tecnico-sperimentale-naturalistico, ma si può estendere anche alle aree storico-umanistiche ed economiche. E ciò è tanto più particolarmente riscontrabile, ancorché oggi in maniera più contenuta, in "arte e scienza" ovvero nell'ambito degli studi nel settore dei beni culturali e ambientali.