DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

The Democratization of Evaluation and Altmetrics

Cinzia Daraio

The Democratization of Evaluation and Altmetrics[1]

Cinzia Daraio

E-mail: daraio@dis.uniroma1.it,
Department of Computer, Control and Management Engineering "Antonio Ruberti" (DIAG), Sapienza University of Rome, Rome (Italy)

Abstract

In this paper we provide an interpretation of the current trends in the evaluation of research and propose a performance evaluation approach for the current time. By adopting an economics and management perspective, we show that there is a need for *democratization* and we propose a possible meaning of Altmetrics as an answer to this need, created -among other factors- by the *unfairness* of traditional bibliometric indicators. The proposed performance evaluation model might be used as a "democratic evaluation tool" for *value creation* in a learning and participatory environment where evaluation is *in* the communities and *for* the communities. Some *caveats* and critical conditions including the relationship between democratization and meritocracy are outlined.

Keywords: research evaluation, Altmetrics, doubly conditional performance evaluation, conditional directional distances, responsible metrics, research integrity

Dated: 15 January 2018

1. Introduction

The evaluation[2] of research activities is a complex task for many reasons. There are no perfect indicators which fit for all purposes. In order to understand the *appropriateness* of the indicators to be used, we need to frame the problem taking into account the systemic nature of the phenomena and to develop models of metrics that are as close as possible to the reality being assessed.

Each metric of research assessment (in this paper metrics and indicators are used as synonym) is based on a *model* that can be implicitly or explicitly defined and discussed. If the model underlying the metric is not described, this does not mean that the indicator is more robust to modelling choice. It simply means that you do not explicitly clarify and account for the underlying theoretical choices, methodological assumptions and data limits. Thus, as a consequence, if you do not specify your model of the metric, you cannot check its *robustness*.

Developing models is important for two main reasons: i) to learn about the explicit consequences of assumptions, test the assumptions, highlight relevant relations; ii) to improve, to better operate, document/verify the assumptions, decompose analysis and synthesis, systematize the problem and the evaluation/choice made, make explicit the dependence of the choice on the scenario. There are however several pitfalls and difficulties in modelling. These mainly relate to the possibility that the targets are not quantifiable; the complexity, uncertainty and changeability of the environment in which the controlled system works and; the limits in the decision context; the intrinsic complexity of calculation of the objective of the analysis. Within this context, Daraio (2017b) proposes a *framework*, which is suggested as a *reference* to develop models of metrics. This paper continues that line of research.

---

[1] This work is based on a paper accepted for presentation and published in the proceedings of the ISSI 2017 Conference (see Daraio 2017a) and a paper presented at the STI/ENID Conference 2017. In Daraio (2017a) the performance evaluation approach was introduced while the discussion about the democratization reported in this paper is new and is an extended and revised version of the one presented at the STI/ENID conference.

[2] In this paper, evaluation is used as synonym of assessment and refers to the different activities related to the assessment process, including setting criteria and formation of judgments.

In this paper, as in the previous one, we do not enter into the discussion about whether it is good or bad to evaluate research, but we adopt a pragmatic view and try to understand the reality and propose a family of models for the evaluation of research, which could fit to the current needs of the field.

The next section describes the current situation, and existing trends, in the evaluation of research. By applying an economic approach, the following section proposes an interpretation of this situation, to identify a trend towards the democratization of the evaluation and it offers a possible meaning to Altmetrics. Next, we introduce our doubly conditional performance evaluation model and explain why it may fit the current needs of the field. The final section outlines the open issues, the caveats and the critical conditions for the implementation of a democratic evaluation tool, including a discussion on the relationship between democracy and meritocracy.

2. Current *Status* and Trends in Research Evaluation

An extensive, complete and current discussion of the current situation in the field of research evaluation would require much more space and is carried out in several recent books (see eg Moed, 2017). In this paper we focus on the main problems and themes, which we summarize, without claiming to be complete, in Table 1.

The assessment of the research activity is also complicated by the *quantification* of data and its processing for use in different contexts and for different purposes (Daraio and Glänzel, 2016), including process monitoring, input-output monitoring, ex-ante and ex-post evaluation. In this case, there is a need to specify *standards* and *rules* for metadata definition and quantification (for additional details and references, see Daraio and Glänzel, 2016). It seems that the need of "a clear and unambiguous terminology and specific standards" (Glänzel, 1996, p.176) is still relevant and timely today.

Table 1. Current *Status* and main Trends in Research Evaluation

Building bock	Selected References and/or main concepts
1. Changes in the production of knowledge	The new production of knowledge. Knowledge and the public in an age of uncertainty (Gibbons et al., 1994; Nowotny et al. 2001)
2. Complexity of the assessment of research	Need to adopt a systematic view, complexity of the assessment linked to the "the implementation problem" (Daraio, 2017a); multidimensionality of the assessment of the research (Moed and Halevi, 2015); problems of data quantification, harmonization and standardization for different evaluation and assessment purposes (Glänzel, 1996, Daraio and Glänzel, 2016; Glänzel and Willems, 2016)
3.Extension to societal values and value for money	Evaluation society (Gläser and Whitley, 2007; Dahler-Larsen, 2011).
4. Introduction of performance-based funding and request of new indicators from policy makers	Greater attention to efficiency and effectiveness of publicly funded research (Hicks, 2009; Jonkers and Zacharewicz, 2016), policy makers increasingly demanding in terms of granularity and cross-referencing of indicators (Daraio and Bonaccorsi, 2017)
5. Rankings and international competition	Proliferation of rankings in a globalized competitive research space
6. Increase of data availability and open access repository	Increase in globally stored information (Hilbert and López, 2011). In the last years there has been an extraordinary development of *open access repositories* all over the world (Pinfield et al. 2014).
7. Introduction and development of Altmetrics	Priem et al. (2010, 2012), Thelwall et al. (2013), Glänzel, and Gorraiz (2015), Haustein, S., Bowman, T. D., & Costas, R. (2015), Costas, C. R., Haustein, S., Zahedi, Z., & Larivière, V. (2016), Ràfols, I., Robinson-García, N., & van Leeuwen, T. (2017)

8. Development of internal research assessment tools	More and more institutions implement internal research assessment processes and build research information systems.
9. Growing of "desktop bibliometrics" 10. Recent critiques of traditional bibliometric indicators	Katz and Hicks (1997)…Publish or Perish, Google Scholar Citation and other commercial products such as Sci-Val and InCites, Cronin and Sugimoto (2014, 2015); Dora Declaration, Leiden Manifesto (Hicks et al., 2015); Wilsdon (2015), Zitt, 2015; Gingras, 2016; Benedictus, Miedema, and Ferguson, (2016), Stephan, Veugelers and Wang (2017)

3. An Interpretation: Towards the Democratization of Evaluation and a Possible Meaning for Altmetrics

To interpret the current situation in the evaluation of research, in this context we apply the economic framework proposed by Acemoglu and Robinson (2006) for analysing the creation and consolidation of *democracy*.

Bunnin and Yu (2004) define democracy as: *"[from Greek demos , people + kratia , mighty, powerful, literally, rule by the people] A form of government, traditionally contrasted to aristocracy (rule by the best), oligarchy (rule by the few), and monarchy (rule by the one). Ideally, democracy requires all citizens to join in making governmental decisions, but such pure democracy, excluding women and slaves, was only practiced for a short period in ancient Athens. The standard democratic form is representative democracy, that is, rule by a group of representatives who are elected for limited periods directly or indirectly by the people. A representative democracy governs through discussion and persuasion rather than by force. Decisions are generally made by majority vote in order that policies will reflect at least to some degree the will or interests of the people. In order to prevent the over-concentration of power, the main legislative, executive, and judicial functions of government are separated. The values and principles underlying this form of government are liberty and equality, sometimes called the democratic ideals."*

For *democratization* we mean "the introduction of a democratic system or democratic principles (https://en.oxforddictionaries.com/definition/democratization, last accessed 10th January 2018)".

Transferred to the field of research evaluation, in this paper, *democratic principles* mean a transparent and *participatory* evaluation system (deliberative policy learning (van den Hove, 2007, Kowarsch, 2016) and *equality* of citizens ("distributive justice" consists in "giving each one his own", see Cozzens 2007 for the concept in the STI policy).

According to Acemoglu and Robinson (2006)[3], the main conditions to create and consolidate democracy are the six reported in the left column of Table 2.

Table 2. An application of the economics of democracy (Acemoglu and Robinson, 2006) to the evaluation of research

Acemoglu and Robinson (2006) component	*Application in the context of research evaluation*
1. *the strength of civil society*	the movement against the *blinkered* use of bibliometric indicators (see point 10 of Table 1)
2. *the structure of political institutions*	-Science-policy interfaces (van den Hove, 2007); -deliberative policy learning (Kowarsch et al. 2016)
3. *the nature of political and economic crises*	the *crisis* of science (see below)

[3] Acemoglu and Robinson (2006) state: "Dictatorship, nevertheless, is not stable when citizens can threaten social disorder and revolution. In response, when the costs of repression are sufficiently high and promises of concessions are not credible, elites may be forced to create democracy. By democratizing, elites credibly transfer political power to the citizens, ensuring social stability. "Democracy consolidates when elites do not have a strong incentive to overthrow it".

4. *the level of economic inequality*	*inequality* which comes out from the skewness of the bibliometric indicators (see below)
5. *the structure of the economy*	structure of the sciences and their linkages: Examples of call for democratization: Hill (2016): "making impact assessment mainstream".
6. *the form and extent of globalization*	form and extent of globalization of science: Examples of call for democratization: Douglass (2016): *The New Flagship University: Changing the Paradigm from Global Ranking to National Relevancy*. Paradeise and Thoenig (2015), unsustainability of the *top of the pile* model.

Let us discuss some of the points identified in the right column of Table 2.

The present *crisis of science* is well summarized in Benessia et al. (2016), who identify the most heated points of discussion in: reproducibility (see Munafò et al. 2017), peer-review, publication metrics, scientific leadership, scientific integrity and the use of science for policy (see also "The end of the Cartesian dream" in Saltelli and Funtowicz, 2015).

van den Hove (2007) defines *science-policy interfaces* "as social processes which encompass relations between scientists and other actors in the policy process, and which allow for exchanges, co-evolution and joint construction of knowledge with the aim of enriching decision-making". van den Hove (2007, p. 824, see also Table 2 p. 815) identifies the following methodological issues to account for in the design, implementation and assessment of the *science-policy* interfaces:

- (i) "the reinforcement and enlargement of scientific *quality and validation processes*;
- (ii) the development of *transdisciplinary* research methodologies;
- (iii) *transparency*, *participation* and *dynamism* of interfaces, in particular the role of other stakeholders and the public;
- (iv) *accountability* of the different actors;
- (v) *translation* of scientific knowledge into policy-relevant knowledge and of policy knowledge into science-relevant knowledge;
- (vi) the inclusion of a *diversity* of knowledges and intelligences;
- (vii) the development of dialogical *dissemination channels* for scientific knowledge which specifically target the various potential user groups; and
- (viii) the *institutionalisation* of science-policy interfaces in a *democratic* context"[4].

Inequality in the scientific production is a well known stylized fact (de Solla Price, 1963; Allison et al., 1982, Allison and Stewart, 1974, Allison, 1980) which is linked to the skewness of the bibliometric indicators (Seglen, 1992; Albarrán et al. 2011; Ruiz-Castillo and Costas, 2014). Recently the issue has been re-actualized by Lok (2016). See also Rousseau and Rousseau (2017).

Nevertheless, in our opinion, the topic of the skewness of bibliometric indicators has not been fully discussed in the field of research evaluation.

We believe it highlights the inequality among scholars and institutions. Scholars in general, even those (the minority) with high performance indicators, think that bibliometric indicators should be handled with care in research assessment. The majority of scholars, which belong to the long queue below the

---

[4] With respect to the last point of van den Hove (the institutionalisation of science-policy interfaces in a democratic context) Kowarsch et al. (2016) identify four main building blocks of deliberative policy learning where deliberative is defined as "inclusive and argumentative way of designing the process" and policy learning as "updating of beliefs about policies resulting from a combination of social interaction, personal experiences, value change and scientific policy analysis". The four blocks are (Kowarsch et al. 2016, p.8 Table 3): i) representation (incorporating wide variety of viewpoints and stakeholders); ii) empowerment (critically scrutinizing requirements to adequately participate; iii) capacity building (building internal capacity of participants: knowledge integration and synthesis; and external capacity building providing knowledge about implications of alternatives, disclosing key uncertainty and normative assumptions); iv) Spaces for deliberation (realizing vertical and horizontal linkages).

average, consider traditional bibliometric indicators as *unfair* or *unjust* tools for research assessment. Indeed, they are so, by construction (see Figure 1).
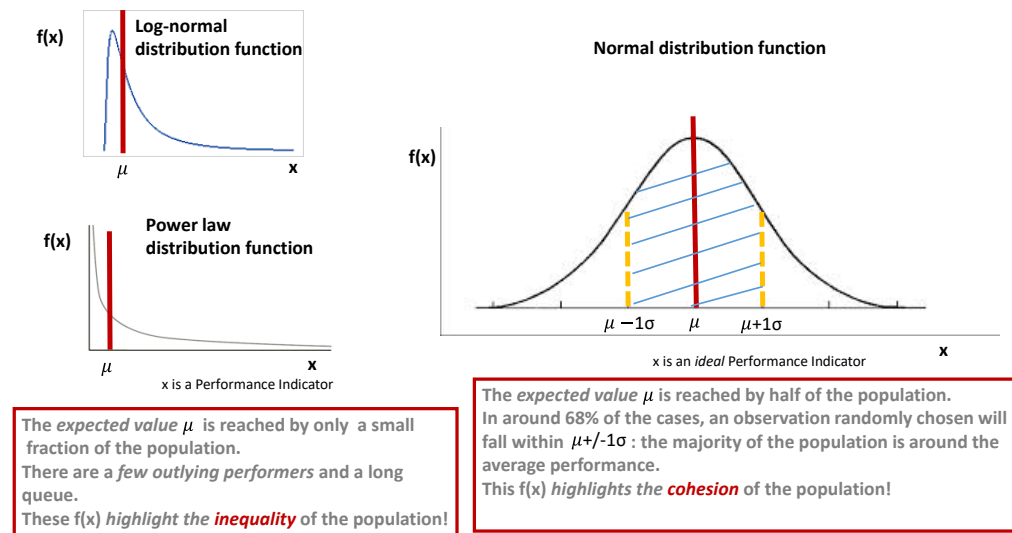
Figure 1. An illustration of the "unfairness" generated by the skewness of performance indicators.

Altmetrics (Priem et al. 2010, 2012) since their origin have been introduced as "alternative measures" to traditional bibliometric indicators, aiming at capturing other impacts dimensions of scholarship activities. They are related to the development of web-based activities. Moed (2017 p.33) links Altmetrics to i) an increasing awareness of the multidimensionality of research performance by policy makers; ii) developments in ICT and social media technologies; iii) Open science movement in the scholarly community.

Although there is an intense research on these metrics (Thelwall et al. 2013; Glänzel, and Gorraiz, 2015; Haustein, Bowman, Costas, 2015; Costas, Haustein, Zahedi, Larivière, 2016; Ràfols, Robinson-García, van Leeuwen, 2017), after more than seven years from their introduction, we do not yet have a clear understanding of what they actually measure and in particular about why scholars decide to commit to altmetrics activities (Wouters, Zahedi and Costas, 2018).

An interpretation that we propose here, supported by the application of the economic of democracy framework carried out in Table 2, is a conjecture that Altmetrics may be interpreted as an answer to the *need for democratization* we detected in the field. This conjecture is supported by the interpretation of Moed (2016) which identifies a link with Nielsen (2012): *the increase of cognitive intelligence could be achieved by conversational critical mass and collaboration which becomes self-stimulating with online tools, which may establish architecture of attention that directs each participant where it is best suited. This collaboration may follow the patterns of open source software: commitment to working in modular way; encouraging small contributions; allowing easy reuse of earlier work; using signalling mechanisms (e.g., scores) to help people to decide where to direct attention.*[5]

---

[5] To highlight the connection of Nielsen (2012) with the democratic values we report another definition of democracy: "… the term "democracy" refers very generally to a method of group decision making characterized by a kind of equality among the participants at an essential stage of the collective decision making. Four aspects of this definition should be noted. First, democracy concerns collective decision making, by which I mean decisions that are made for groups and that are binding on all the members of the group. Second, this definition means to cover a lot of different kinds of groups that may be called democratic. So there can be democracy in families, voluntary organizations, economic firms, as well as states and transnational and global organizations. Third, the definition is not intended to carry any normative weight to it. It is quite compatible with this definition of democracy that it is not desirable to have democracy in some particular context. So the definition of democracy does not settle any normative questions. Fourth, the equality required by the definition of democracy may be more or less deep. It may be the mere formal equality of one-person one-vote in an election for representatives to an assembly where there is competition among candidates for the position. Or it may be more robust, including equality in the processes of deliberation and coalition building. "Democracy" may refer to any of these political arrangements. It may involve direct participation of the members of a society in deciding on the laws and policies

This is particularly useful to connect Altmetrics, its connection with the Open Science movement, with Van den Hove (2007), Ewell (2009) and Kowarsch et al. (2016). Our interpretation is also coherent with the name given by the authors who introduced it, namely *alternative* metrics that is metrics which are *substitutive* of the traditional ones, *not complimentary* ones (for a critique of the name, altmetrics as a "good idea, but with a bad name" see Rousseau and Ye, 2013).

In the next section, after an *excursus* on the performance assessment in economics and business, we propose a performance evaluation model that may be used as *a tool for the democratization of evaluation*: it may permit the *co-existence* of heterogeneous performance indicators. It could be a first step towards a formative democratic approach to evaluation in which indicators are used as learning tools instead of target of policy.

4. A Performance Evaluation Approach for Current Time

4.1 Assessment of the performance: what can be learned from the business sector?

In the evaluation of research, we have seen a need of changing from a traditional performance evaluation, based on traditional bibliometric indicators of number of publications and citations, towards a multidimensional performance model which includes alternative and impacts metrics. For many years, a similar pattern has been observed, in business performance measurement. As reconstructed in Ghalayini and Noble (1996), the business performance measurement has two phases. In the first phase (from 1880s to 1980s) the emphasis was on financial measures such as profit, return on investment, and productivity. The second phase started in the late 1980's as a result of changes in the world market. To regain a competitive edge, companies shifted their strategic priorities from low-cost production to quality, flexibility but also implemented new philosophies of production management (Computer Integrated Manufacturing, Just in Time, Total Quality Management and so on). These changes have revealed that there is a need to develop new performance measurement systems. As the Performance Measurement Manifesto (Eccles, 1991, p. 131) states: "At the heart of this revolution lies a radical decision: to shift from treating financial figures as the foundation for performance measurement to treating them as one among a broader set of measures". In the Performance Manifesto the revolution is all about combining information systems and human resources (called as a culture shock).

In this literature (see e.g. Simons, 2000; Neely, 2002) there has been an evolution of the performance evaluation from the classical "shareholder" value approach (based on the maximization of profits) to the balanced score-card approach (Kaplan and Norton, 1992) which extends the performance measurements to 4 dimensions. In addition to the classical economic perspective, they include also the other perspectives of internal process, of customers, and finally of the growth and learning. recent examples of multi-dimensional performance evaluation models are the multi-stakeholder model by Atkinson, Waterhouse and Wells (1997) and the performance prism model by Neely, Adams and Kennerley (2002), which are. The most recent trends in this sector see the inclusion of indicators of *sustainability* and the development of *social corporate responsibility* awareness.

The discussions and developments of the business performance measurement literature might be helpful, for the "modernization" of the performance evaluation in the assessment of research. In particular, *mutatis mutandis*, the aspects of the performance measurement system design (see e.g. Neely, Gregory and Platts, 1995) and the implementing and updating aspects of the performance measurement systems (see e.g. Bourne et al. 2000), may be interesting to consider.

4.2 Recent trends in the New Public Management

Performance measurement in public management is a challenged subject (Johnsen, 2005). Woelert (2015) identifies a proliferation of performance indicators in a kind of technical and quantitative escalation. Lewis (2015) highlights the almost neglected consideration of the consequences of performance measurement in evaluation of public policies.

---

of the society or it may involve the participation of those members in selecting representatives to make the decisions."(Tom, 2015).

The principles of performance measurement of the business sector have been extended to the evaluation of the public administration sectors, giving rise to the birth of New Public Management (see e.g. Ferlie,1996; Lane, 2000; Kaplan, 2001). Recently, we can observe a trend towards the "democratization" in this context too (see e.g. Behn, 1998; Box et al. 2001) so that it has been described as a paradigm shift (Dunleavy et al. 2006, O'Flynn, 2007) and linked to the need of "Citizen-driven" performance measurement (see Van Dooren, Bouckaert and Halligan, 2015).

4.3 Assessment of the performance from an econometric perspective: productivity and efficiency analysis

In economics, the theme of performance measurement is very old, dating back to Adam Smith's pin factory if not before. Traditionally, performance evaluation, intended as *productivity*, which is how the output is produced taking into account the inputs used, was performed based on the specification of a "production function" that described the relationship between the resources used (production factors or inputs) and the output produced (the result obtained from the production process). Implicit in this definition was a hypothesis of maximization of agents' behaviour, as the production function represented the maximum output achievable within the specified production set, for certain levels of input used. Often this type of production function was (and still is) estimated through linear regressions, with the method of least squares (OLS) that capture the average performance trend. Subsequently, starting from the works of Farrell (1957) based on the works of Koopmans and Debreu (see also Shephard, 1970) a non-mainstream approach was introduced to the measurement of productivity. This approach does not assume the hypothesis of efficiency as implicit, but it estimates its level. It is the production frontier approach (see Färe, Grosskopf and Lovell, 1994).

Production functions aim at capturing, through the estimated coefficients, an "average" productive behaviour that is a kind of "representative" behaviour of the units analysed. Production frontiers, instead, extend this framework introducing the idea of "efficiency" as a distance of productive units from an estimated efficient frontier or frontier of the best practices. In this framework, efficiency is measured as the maximal achievable output with respect to an "efficient frontier" estimated on the data. To estimate an efficient frontier, both parametric and nonparametric approaches have been developed (see Bonaccorsi and Daraio, 2004 for an overview on the econometric approaches to assess research productivity).

According to this paper, the most interesting approach for the assessment of research activities is the nonparametric one. This is because nonparametric methods, as their name states, are approaches that do not rely on the specification of parameters to characterize the relations among the entities under analysis. Generally, as observed by Stock (2010), nonparametric approaches in econometrics have been developed thanks to the development of computer power and the advancements of mathematical and statistical research. Stock (2010) identifies one of the causes of the development of nonparametric models in dissatisfaction towards traditional parametric models, which were not always a good approximation of complex economic realities.

Indeed, as has been observed in many contexts of the public sector domains, where there are multiple inputs, multiple outputs realities and not available price information, the nonparametric approach, relying on a minimal set of assumptions, represents a very good approach to the quantitative performance assessment (Thanassoulis et al, 2008; Zhu, 2014).

It is interesting to note that these literatures (the one on the business evaluation and the economic one) have developed in parallel, with few points of contact up to recent years. Only recently (see, for example Chen, Delmas and Lieberman, 2015) have there been connections between the two. In particular, some key concepts from the efficiency analysis (including *directional distances,* Chambers et al. 1998, Fare and Grosskopf, 2000) are starting to be used for the strategic performance evaluation (Bogetoft, 2012) in a performance management perspective. Efficiency analysis techniques are able to overcome some of the main limitations of traditional key performance indicators, namely their fragmentation, providing an overall quantitative assessment.

The performance evaluation model we propose in this paper may be implemented by applying recently developed tools from the nonparametric approach to efficiency analysis. We refer in particular to conditional directional distance models (Daraio and Simar, 2014, 2016) which allow a great flexibility

in terms of specification of the path to reach the efficient frontier (in terms of input (x) or output (y) direction setting) as well as allow us to account for the heterogeneity of the analysed units, including conditioning external or environmental factors (called Z in the literature). At the end of the next section, we report an example of possible implementation.

4.4 A Doubly-Conditional performance evaluation model

The output of the research activity has some features that include complexity, uncertainty and indeterminacy. Among the challenges of their assessment, there are: 1) to bring about communication and debate about assumptions, choices and uncertainties, and about the limits of scientific knowledge; 2) to allow for articulation of different types of (scientific, local, indigenous, political, moral and institutional) knowledge; 3) to provide room for a transparent negotiation among standpoints (participatory processes)(van den Hove, 2007, see p. 815 for more normative requirements for science-policy interface).

The performance evaluation model we will introduce in this section can be helpful in combining the advantages of the partial convergence indicators approach (Martin and Irvine, 1983; Martin, 1996) mainly measurability and possibility of calculating the indices of performance, with a need for a multidimensional approach to the assessment of research and its impact (Moed and Halevi, 2015).

Martin and Irvine (1983) introduce the idea of converging *partial* indicators approach for assessing scientific performance, based on the multidimensional nature of research and its outputs[6]. Given the partial character of indicators, it is only possible to draw reliable conclusions in those cases in which the indicators provide convergent results, keeping the influence of non-relevant factors low. Moed and Halevi (2015) on the other hand, propose "a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, based on the notion that indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups." Diverse institutional missions, and different policy environments and objectives, require different assessment processes and indicators. They focus on the purpose, objectives, and policy context of research assessments, and demonstrate how these characteristics determine the methodology and metrics to be applied.

For instance, publication counts are useful instruments to help in discriminating between those staff-members who are research active and those who are not, but are of little value if research active scientists are to be compared one with another according to their research performance. This idea is further discussed as *minimum performance standards* in Moed (2017, p.31). Moed and Halevi (2015) introduce the concept of a *meta-analysis* of the units under assessment in which metrics are not used as tools to evaluate individual units, but to reach policy decisions regarding the overall objective and general setup of an assessment process. Their base underlying assumption, that we share here, is that "the future of research assessment exercises lies in the *intelligent* combination of metrics and peer review. A necessary condition is a thorough awareness of the potentialities and limitations of each method."

Figure 2 illustrates the main components of the *doubly conditional* performance evaluation model we propose here, which are based on a combination and extension of elements from Johnsen (2005); van den Hove (2007) and Lewis (2015). It is "doubly conditional" because the evaluation is conditioned twice: on the information that is available and on that which is not available.

In Figure 2, the different types of arrows and the different types of boxes simply illustrate *heterogeneous* influences and contents respectively. The elements filled with gray represent the main items on which the conditioning could be carried out. By adopting a process based view, the model identifies different objects of the evaluation: outputs (the results obtained from the combination of the inputs during the process), efficiency (relation between the outputs and the inputs), effectiveness (the efficiency with respect to the objectives) and finally impacts.

---

[6] "All quantitative measures of research are, at best, only partial indicators influenced partly by the magnitude of the contribution to scientific progress and partly by other factors. Nevertheless, selective and careful use of such indicators is surely better than none at all. Furthermore, the most fruitful approach is likely to involve the combined use of multiple indicators. However, because each is influenced by a number of other factors, one needs to try and control for those by matching the groups to be compared and assessed as closely as one can" (Martin, 1996, p. 351).

We distinguish two kinds of conditioning. Conditioning on the items reported in the bottom of the Figure 2 (policy, objectives, actors, processes and results) means to compare comparable entities, setting appropriate reference sets. We call this *internal conditioning* or normalization. On the other hand, conditioning on the items reported in the top of the Figure 2 accounts for heterogeneity factors that we call *external conditioning* or contextualization. According to this model of performance evaluation, it is all a matter of appropriate normalization and contextualization.

This model:

1. Permits the *identification* of the *components* of the analysis (in terms of model characterization) that are excluded (what remains outside) in the specific context of the evaluation;

2. Gives *interpretative value* to the metrics calculated. The metrics have to be considered as *residuals*, i.e. what remains after the consideration of the dimensions we pursued, that is due to other factors/components not accounted for;

3. Represents a step toward the *democratization* of the evaluation practice, able to balance the opposite views of external accountability and internal improvement (Ewell, 2009), composing contrasting trends towards *competition* and *cooperation* through cohesion.

This performance evaluation model might be helpful to identify *constitutive effects* of indicators (Dahler-Larsen, 2014) and perhaps also their "unintended consequences". This model is useful for the interpretation of the results obtained from our assessment, and it is useful to identify discrepancies, and shows what the *residual*, our "ignorance" is. It is also helpful to identify the gap and which variables may be added to explain a part of these discrepancies. It is a kind of contemporary revisiting and revalidation of the Leibnsteins *x-inefficiency* concept (Leibenstein 1966, 1975, 1978a,b; see also Leibenstein and Maital, 1992).
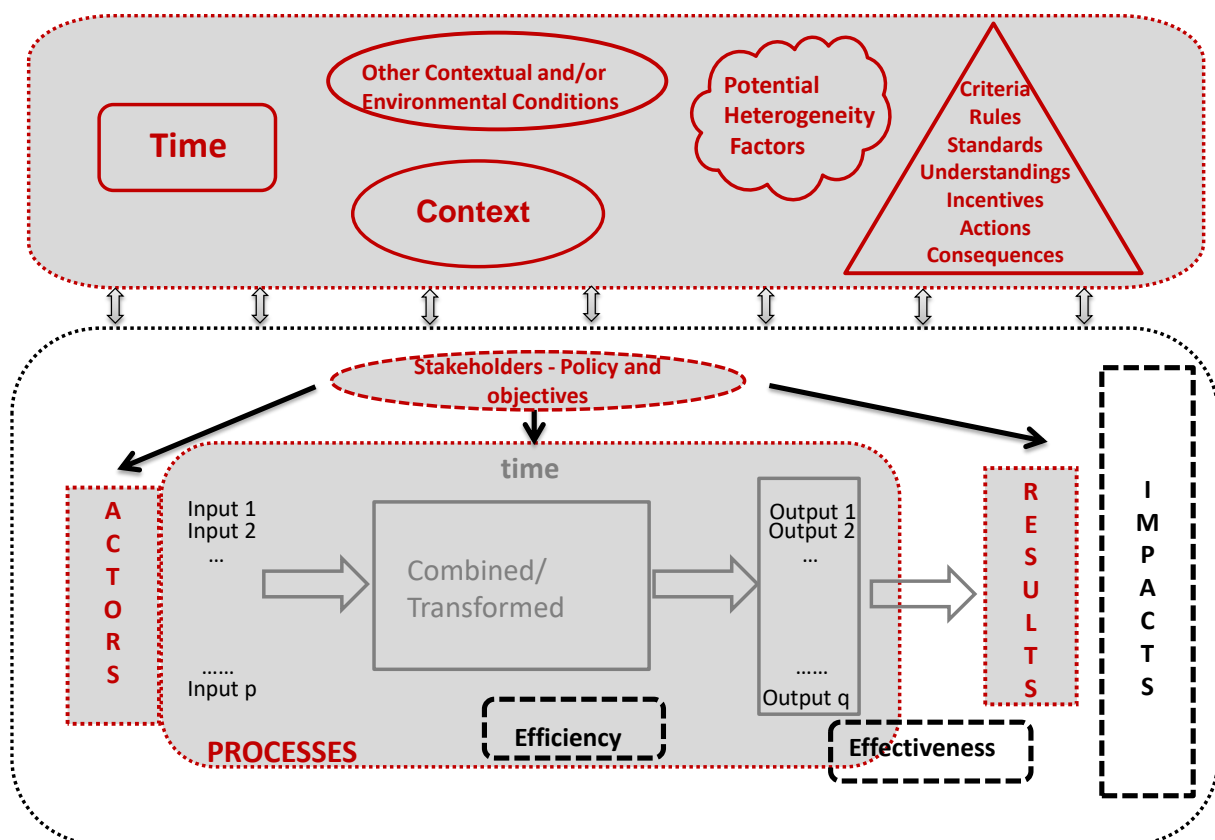

Figure 2. A "doubly conditional" performance evaluation model

This doubly conditional performance evaluation model may be used for *profiling* the activities, sharing evaluation models, improving knowledge by learning. This *performance evaluation model* may be used as a *flexible evaluation tool* in a learning and participatory environment. It might be seen as a revisited version of Ricardo's approach of *comparative advantages* (Ricardo, 1817) but within the context of a broader framework. Within this context, *for each subject under assessment a dimension of performance along which the evaluated entity can outperform or do better than the others can be*

found. The identification of the best performing dimension of each entity subjected to the evaluation is important for: developing strategy to identify and establish *sustainable* and durable *value creation*, going *beyond competitive advantages* (Zenger, 2016), exploiting the existing *information* (Porter and Millar, 1985). Each subject under evaluation would be able to find out their own *specialty-role* in the knowledge production system. This way, the evaluation is done *in* the communities and *for* the communities.

4.5 An illustrative example

An example of application of the proposed doubly conditional performance model described so far may be the evaluation of the scholars of a department of a generalist university. In order to implement the approach it is necessary to design and develop a shared and distributed *information system*, possibly open to add new dimensions, including: publications (different outputs), citations, teaching hours, courses, administrative duties, research projects, contract research, third mission activities, altmetrics,…). The information system should be accessible to the individuals, at the department and university governance. For the assessment of the Ricardian measures of comparative advantages we may apply the Conditional Directional distance measures (Daraio and Simar, 2014, 2016). Conditional directional measures are based on directional distance measures, which account for the *contextual* variables and hence handle heterogeneity by reducing the comparison set to include only those that share a similar context. Figure 3 illustrates a simple directional distance.
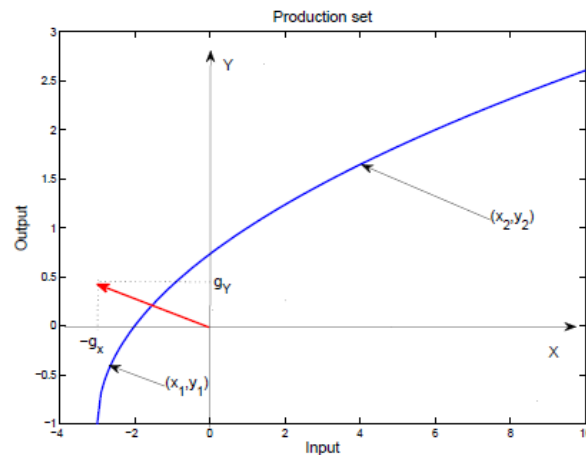
Figure 3. An illustration of a directional distance.
From a technical point of view, a directional distance projects a unit, represented by a vector of inputs and outputs (x,y), onto an efficient frontier estimated over a sample of data along the direction $g=(-g_x,g_y)$. A performance measurement d is calculated as the distance of the unit from the frontier:
$$d(x,y|g_x,g_y, \Psi)=\sup\{d|(x-dg_x,y+dg_y) \in \Psi\},$$ where Ψ is the production set.

Described in a simple way, the conditional directional approach tells us in this context that if we account for all possible inputs, outputs and external conditioning factors all the entities under evaluation become efficient, reach their own conditional efficient frontier. It is then a matter of selecting the direction through which reach the best practice frontier on the base of the *relative comparative advantages*. Provided that minimum thresholds in some relevant dimensions (for instance, professors in Italy have the duty by law to teach and doing research, so minimum thresholds for these dimensions) may be added, the subjects of the assessment are *free* to choose the best combination of activities to carry out. This implies that scholars, department and university governance all components share the information system, and can use it to check the underlying data, assess the performance, their progress and so on. It could help to develop participation, interaction, and learning aspects of the units analyzed in a multidimensional evaluation framework where each one can find its own best way to contribute to the "creation of value" of its over-hierarchical institution (for example, for individuals the department of the university, for the university the region in which it is located, and so on).

5. Conclusions, Caveats, Open Issues and Further Research

This paper offers an interpretation of the existing current trends in the evaluation of research. From our analysis there seems to be a trend towards the *democratization* of evaluation, which is the need to introduce democratic principles characterized by social equality, representativeness, transparency, partecipation, open deliberation. We propose an interpretation of Altmetrics as an answer to this need of democratization. We introduce a *doubly conditional performance evaluation model* which can be used as a *democratic evaluation tool* for "value creation" in a learning and participatory environment. According to this performance evaluation model, if we are able to account for all its components, the entities under assessment (e.g. scholars, institutions...), may be seen as performing at the best of their possibilities. Put in another way, *for each entity there is an "optimal path" or trajectory in the multidimensional performance framework along which it is top performing, which permits to reach its best possible result.* The *democratization* of evaluation relies on the balance between external accountability and internal improvement (strategic management and decision-making) in which indicators are used as learning tools instead of target of policy.

The findings of this paper seem to show that perhaps, the critique of traditional bibliometric indicators (constructed on number of publications and citations) is exacerbated by some *unpleasant* and *tricky* properties these indicators have, e.g. skewness and asymmetry, which highlight the *inequality* among the assessed units. The critiques of bibliometric indicators have increased over the years, also because, among other factors, there has been an increasing usage of bibliometric indicators at the individual level. It is well known that at aggregate level, some problems of bibliometric indicators compensate and then these indicators can be valid tools to investigate for instance the dynamics of disciplinary profiles of countries. Nevertheless, when indicators are used in research assessment in which individuals are the unit of the analysis, *much more care* should be given to the aspects of "democratization".

Democracy, on the other hand, is a delicate word that evokes emotional and philosophical responses. Someone may suggest that legitimation of evaluation would be better than democratization of evaluation. Some may agree with our interpretation, because they like the open evaluation idea, the idea of co-creation of value, the mixing of a producer and a consumer approach in this context.

On the other hand, democracy makes people think about *equality*, and someone may think it is not appropriate to the research activity, a field in which we do not have homogenous intelligences and talents. According to this perspective, an assessment should find out the "best", should select those that *merit* to be selected, not the elected by the majority of the population. Now we come to a tricky issue, which is the relationship between *democracy* and *meritocracy*. More specifically, if it is possible to account for meritocracy in a democratic evaluation context.

Young (1958) defines merit as the sum of intelligence and effort. Nevertheless, one of the primary concerns with meritocracy is the ambiguous (unclear) definition of "merit" (Arrow et al. 1999). Recent studies (Sternberg and Kaufman, 2011; Kaufman, 2013) show that "greatness" is more than just the sum of the "nature" and "nurture" components, and to understand it we have to go beyond talent and practice. Carson (2007) in *The Measure of Merit* shows that talents and intelligence have become constituents of the societies in which they were produced and adopted, continually shaping and being shaped by these cultures. The concepts of intelligence and merit, hence, remain always contestable terms in the recurrent debates about the social and political implications of inequality for a modern democracy (Carson, 2007). In another paper (Ruocco, Daraio, Folli and Leonetti, 2017), we have shown the ambiguity of bibliometric indicators in relation to unobserved (unspecified) abilities.

This opens the reflection on which evaluation model to adopt, or better, which kind of models -families of models- to adopt because due to the heterogeneity and the diverse purposes and ways of doing evaluation, of course there is not *one best way*. A unique model which fits for all does not exist; a model may be good for a given purpose but not appropriate for another, the "quality" of the model should be defined as "fitness for purpose" of evaluation.

This is a relevant question because it is linked to the future of research evaluation. Given that the models of evaluation have implications and change the behaviour of people that are evaluated, this question also has implications on the research activities themself.

Which model of democracy, which level of democracy, what are the best "open channels" for representativeness, citizenship and participations are all open questions.

The consideration and implementation of the main *substantial* and *formal* criteria of democracy - including division of powers, no conflict of interests, decentralization, contextualization- are other all open questions to address. They include issues related to the data platforms, which solutions, if private or publicly owned and so on.

In this paper, given our background, we have only *superficially* started to address the issue of democratization. In our opinion, *normative democratic theory* is a serious subject, which would require much more deep and interdisciplinary research in this evaluation context.[7] As described by Tom (2015) it is linked to the *principles* and *conceptions* of *humanity* and *society*.

Interestingly, in the conclusion of a recent report on *Fostering Integrity in Research* (2017, p.208-209) by the US National Accademies, it is stated: "The committee reaffirms the central recommendation from *Responsible Science* [a previous report of 2002]that formally places the *primary responsibility* for acting to define and strengthen basic principles and practices for the responsible conduct of research on *individual scientists* and *research institutions*. At the same time, the committee based its recommendations on its understanding that the integrity of research depends on creating and maintaining a system and environment of research in which institutional arrangements, practices, policies, and incentive structures support responsible conduct. Fostering research integrity is an obligation shared not only by individual researchers but also by leaders and those involved with all organizations sponsoring, conducting, or disseminating research, including corporate and government research organizations."

Hence, the primary responsibility is on individual scientists.

Fraud and misconduct have for several years been identified as a relevant problem of the scientific community. For a review, see Fanelli (2009). In the recent years, with the advent of the evaluation society, the situation has received an increasing attention, up to the introduction and discussion of a retraction index (Fang, and Casadevall, 2011). More recently, Fang et al. (2012) found that the main reason for retractions relies on misconduct.

We could learn more about evaluation, going deeper and investigating the human behaviour (e.g. Arendt, 1958). This may require an "integral" approach to the quantitative evaluation of science, which includes philosophy and history.

Acknowledgments

---

[7] A definition of the function of normative democratic theory is: "The function of normative democratic theory is not to settle questions of definition but to determine which, if any, of the forms democracy may take are morally desirable and when and how. For instance, Joseph Schumpeter argues (1956, chap. XXI), with some force, that only a highly formal kind of democracy in which citizens vote in an electoral process for the purpose of selecting competing elites is highly desirable while a conception of democracy that draws on a more ambitious conception of equality is dangerous. On the other hand, Jean-Jacques Rousseau (1762, Book II, chap. 1) is apt to argue that the formal variety of democracy is akin to slavery while only robustly egalitarian democracies have political legitimacy. Others have argued that democracy is not desirable at all. To evaluate their arguments we must decide on the merits of the different *principles and conceptions of humanity and society* from which they proceed" (Tom, 2015).

References

Acemoglu, D., & Robinson, J. A. (2006). Economic origins of dictatorship and democracy. Cambridge University Press.

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics, 88(2), 385-397.

Allison, P. D. (1980). Inequality and scientific productivity. Social Studies of Science, 10(2), 163-179.

Allison, P. D., & Stewart, J. A. (1974). Productivity differences among scientists: Evidence for accumulative advantage. American sociological review, 596-606.

Allison, P. D., Long, J. S., & Krauze, T. K. (1982). Cumulative advantage and inequality in science. American Sociological Review, 615-625.

Arendt H. (1958), The Human Condition, 2nd ed. Chicago: University of Chicago Press.

Arrow, Bowles and Durlauf. Meritocracy and Economic Inequality. Princeton, 1999.

Atkinson, A. A., Waterhouse, J. H., & Wells, R. B. (1997). A stakeholder approach to strategic performance measurement. Sloan management review, 38(3), 25.

Behn, R. D. (1998). The new public management paradigm and the search for democratic accountability. International Public Management Journal, 1(2), 131-164.

Benedictus, R., Miedema, F., & Ferguson, M. W. (2016). Fewer numbers, better science. Nature, 538(7626).

Benessia, A., Funtowicz, S., Giampietro, M., Pereira, A. G., Ravetz, J., Saltelli, A., ... van der Sluijs, J. P. (2016). Science on the Verge. Amazon Book, in the Series The Rightful Place of Science Consortium for Science, Policy Outcomes Tempe, AZ and Washington, DC.

Bourne, M., Mills, J., Wilcox, M., Neely, A., & Platts, K. (2000). Designing, implementing and updating performance measurement systems. International journal of operations & production management, 20(7), 754-771.

Box, R. C., Marshall, G. S., Reed, B. J., & Reed, C. M. (2001). New public management and substantive democracy. Public Administration Review, 61(5), 608-619.

Bunnin N. and Yu J. (2004), Democracy, in The Blackwell Dictionary of Western Philosophy, eISBN: 9781405106795, last accessed 7th January, 2018.

Costas, C. R., Haustein, S., Zahedi, Z., & Larivière, V. (2016). Exploring paths for the normalization of altmetrics: Applying the Characteristic Scores and Scales. The 2016 Altmetrics Workshop, Bucharest, Romania.

Cozzens, S. E. (2007). Distributive justice in science and technology policy. Science and Public Policy, 34(2), 85-94.

Cronin, B., & Sugimoto, C. R. (2014). Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact. MIT Press.

Cronin, B., & Sugimoto, C. R. (Eds.). (2015). Scholarly metrics under the microscope: from citation analysis to academic auditing. Information Today, Medford USA.

Dahler-Larsen, P. (2011). *The evaluation society*. Stanford University Press.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: getting beyond unintended consequences. *Public Management Review*, 16(7), 969-986.

Daraio C. (2017a), Assessing research and its impacts: The generalized implementation problem and a doubly-conditional performance evaluation model, *ISSI 2017 - 16th International Conference on Scientometrics and Informetrics, Conference Proceedings*, pp. 1546-1557.

Daraio C. (2017b). A Framework for the Assessment of Research and its Impacts. *Journal of Data and Information Science,* Vol. 2 No. 4, 2017 pp 7–42.

Daraio C., Glänzel W. (2016), Grand Challenges in Data Integration. State of the Art and Future Perspectives: An Introduction, Scientometrics, 108 (1), 391-400.

Daraio C. Simar L. (2014), - Directional Distances and their Robust versions. Computational and Testing Issues, European Journal of Operational Research, 237, 358-369

Daraio, C., Simar L. (2016), Efficiency and benchmarking with directional distances: a data-driven approach, Journal of the Operational Research Society, 67 (7), 928-944

de Solla Price, D. J. (1963). Little science, big science. New York: Columbia University Press.

Douglass, J. A. (Ed.). (2016). The New Flagship University: Changing the Paradigm from Global Ranking to National Relevancy. Springer.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). New public management is dead—long live digital-era governance. Journal of public administration research and theory, 16(3), 467-494.

Eccles, R. (1991). The performance measurement manifesto. Harvard business review, 69(1), 131-137.

Ewell, P. (2009). Assessment, Accountability and Improvement: Revisiting the Tension. National Institute for Learning Outcomes Assessment, http://www.learningoutcomesassessment.org/documents/ PeterEwell_005.pdf, last accessed, 28 September 2016.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS one, 4(5), e5738.

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. Infection and immunity, 79(10), 3855-3859.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences, 109(42), 17028-17033.

Ferlie, E. (1996). The new public management in action. Oxford University Press, USA.

Ghalayini, A. M., & Noble, J. S. (1996). The changing basis of performance measurement. International Journal of Operations & Production Management, 16(8), 63-80.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). The new production of knowledge: The dynamics of science and research in contemporary societies. Sage.

Gingras, Y. (2016). *Bibliometrics and research evaluation: Uses and abuses*. MIT Press.

Glänzel, W. (1996). The need for standards in bibliometric research and technology. Scientometrics, 35(2), 167-176.

Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: confusing terminology?. Scientometrics, 102(3), 2161-2164.

Glänzel, W., Willems, H., (2016), Towards standardisation, harmonisation and integration of data from heterogeneous sources for funding and evaluation purposes. Scientometrics, 106 (2), 821–823.

Gläser, J., & Whitley, R. (Eds.). (2007). The changing governance of the sciences: The advent of research evaluation systems. Springer.

Haustein, S., Bowman, T. D., & Costas, R. (2015). Interpreting" altmetrics": viewing acts on social media through the lens of citation and social theories. arXiv preprint arXiv:1502.05701, to be published in: Cassidy R. Sugimoto (Ed.). *Theories of Informetrics*: A Festschrift in Honor of Blaise Cronin.

Hemlin, S., & Rasmussen, S. B. (2006). The shift in academic quality control. Science, Technology & Human Values, 31(2), 173-198.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature, 520, 429-431.

Hill, S. (2016). Assessing (for) Impact: Future Assessment of the Societal Impact of Research. Palgrave Communications, 2, 16073.

Johnsen, Å. (2005). What does 25 years of experience tell us about the state of performance measurement in public policy and management? Public Money and Management, 25(1), 9-17.

Kaplan R., Norton D., The Balanced Scorecard - Measures that Drive Performance, Harvard Business Review, 1992

Kaplan, R. S. (2001). Strategic performance measurement and management in non-profit organizations. Non-profit management and Leadership, 11(3), 353-370.

Kaufman, S. B. (Ed.). (2013). The complexity of greatness: Beyond talent or practice. Oxford University Press.

Kowarsch, M., Garard, J., Riousset, P., Lenzi, D., Dorsch, M. J., Knopf, B., ... & Edenhofer, O. (2016). Scientific assessments to facilitate deliberative policy learning. Palgrave Communications, 2, 16092.

Lane, J. E. (2000). New public management. Taylor & Francis US.

Lewis, J. M. (2015). The politics and consequences of performance measurement. Policy and Society, 34(1), 1-12.

Liu, A. (2011). Unraveling the myth of meritocracy within the context of US higher education. Higher education, 62(4), 383-397.

Lok, C. (2016). Science's 1%: How income inequality is getting worse in research. Nature News, 537(7621), 471.

Moed, H. F. (2016). Altmetrics as traces of the computerization of the research process. In C. R. Sugimoto (Ed.), Theories of informetrics and scholarly communication. A Festschrift in honor of Blaise Cronin (pp. 360-371). Berlin: De Gruyter.

Moed, H. F. (2017). Applied Evaluative Informetrics. Springer International Publishing.

Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. Journal of the Association for Information Science and Technology, 66(10), 1988-2002.

National Academies of Sciences, Engineering, and Medicine, (2017), Fostering Integrity in Research. Washington, DC: The National Academies Press.

Neely, A. (Ed.). (2002). Business performance measurement: theory and practice. Cambridge University Press.

Neely, A. D., Adams, C., & Kennerley, M. (2002). The performance prism: The scorecard for measuring and managing business success. London: Financial Times/Prentice Hall.

Neely, A., Gregory, M., & Platts, K. (1995). Performance measurement system design: a literature review and research agenda. International journal of operations & production management, 15(4), 80-116.

Nielsen, M. (2012). Reinventing discovery: the new era of networked science. Princeton University Press.

Nowotny, H., Scott, P., Gibbons, M., (2001). Re-thinking science: Knowledge and the public in an age of uncertainty (p. 12). Cambridge: Polity.

O'Flynn, J. (2007). From new public management to public value: Paradigmatic change and managerial implications. Australian journal of public administration, 66(3), 353-366.

Paradeise, C., & Thoenig, J. C. (2015). *In Search of Academic Quality*, Springer International.

Porter, M. E., Millar, V. E. (1985). How information gives you competitive advantage.

Priem J., D. Taraborelli, P. Groth, C. Neylon (2010), Altmetrics: A manifesto, 26 October 2010. http://altmetrics.org/manifesto

Priem, J., Groth, P., & Taraborelli, D. (2012). The altmetrics collection. PloS one, 7(11), e48753.

Ràfols, I., Robinson-García, N., & van Leeuwen, T. (2017). How to make altmetrics useful in societal impact assessments: shifting from citation to interaction approaches. Impact of Social Sciences Blog.

Rousseau R., and Ye Y. F. (2013). "A multi-metric approach for research evaluation." Chinese Science Bulletin 58.26, 3288-3290.

Rousseau S., Rousseau R. (2017), Inequality in science and the possible rise of scientific agents, ISSI Newsletter 12, 4, p. 68-70.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. Journal of Informetrics, 8(4), 917-934.

Ruocco G., Daraio C., Folli V., Leonetti M., (2017), Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent. Palgrave Communications. 3:17064 doi: 10.1057/palcomms.2017.64.

Seglen, P. O. (1992). The skewness of science. Journal of the American Society for Information Science, 43(9), 628.

Simons, R. (2000). Performance measurement and control systems for implementing strategy. Prentice Hall.

Stephan P., Veugelers R. and Wang J. (2017), Blinkered by bibliometrics, Nature, 544, 411-412.

Sternberg, R. J., & Kaufman, S. B. (Eds.). (2011). The Cambridge handbook of intelligence. Cambridge University Press.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PloS one, 8(5), e64841.

Tom C. (2015), "Democracy", The Stanford Encyclopedia of Philosophy (Spring 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2015/entries/ democracy/>. Last accessed 7th January 2018.)

Van den Hove, S. (2007). A rationale for science–policy interfaces. Futures, 39(7), 807-826.

Van Dooren, W., Bouckaert, G., & Halligan, J. (2015). Performance management in the public sector. Routledge.

Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. DOI: 10.13140/RG.2.1.4929.1363.

Wouters P., Zahedi Z., Costas R. (2018), Social media metrics for new scientific evaluations, in Handbook of Science and Technology Indicators, eds., Springer, *forthcoming*.

Young, M. D. (1958). The rise of the meritocracy. Transaction Publisher.

Zenger, T. (2016). Beyond Competitive Advantage. Harvard Business Review Press, Boston.

Zitt, M. (2015). The excesses of research evaluation: The proper use of bibliometrics. *Journal of the Association for Information Science and Technology*, 66(10), 2171–2176.