Designing indicators for opening up evaluation: insights from research assessment

Ràfols, Ismael; Stirling, Andy; Dahler-Larsen, Peter

10 Designing indicators for opening up evaluation: insights from research assessment

Ismael Ràfols and Andy Stirling

Introduction

Over recent decades, indicators have become increasingly prominent in governance across a variety of sectors. Through most of the twentieth century, indicators were used as tools to inform, support, and justify decision making. The advent of neoliberal governance and New Public Management (NPM) in the 1980s brought about increased use of *scalar quantification*. By this, we mean that issues of interest are not just represented quantitatively, but as single (notionally definitive) numbers. Thus, in both private and public spheres, NPM aims to help ensure greater adoption of "best" choices (i.e. more productive, more efficient, higher performance) (Desrosières, 2015). This overall increase is associated with greater use of indicators in micro-management of organizations and staff, underpinning more explicit incentive mechanisms for fostering self-monitoring, self-auditing, and external control towards improved "performance" (Dahler-Larsen, 2011; Rottenburg & Merry, 2015, pp. 3–7). The internet and growing computational capacity further fueled these changes through "metrics" developed from big data analytics.

This expansion and growth of indicators has been seen as problematic. By contrast with more "plural and conditional" numerical "mappings" of views under diverse perspectives, scalar quantifications encode particular understandings and interests on what counts and how it is counted. Underpinning globally burgeoning information infrastructures, such conventional indicators tend to privilege the perspective of those with financial and professional resources (Rottenburg & Merry, 2015, p. 4). Evaluations shaped by scalar indicators are

likely to privilege particular understandings of performance – at the expense of alternatives whose value is not so easily captured by these indicators. Application of this kind of indicator-based evaluation is likely to result in a narrowing of activities, with those not counting for the indicators being suppressed (*task reduction*) and indicators themselves becoming a goal in their own right (*goal displacement*) (de Rijcke, Wouters, Rushforth, Franssen, & Hammarfelt, 2016). Such constitutive effects resulting from indicator-centered evaluation are often perceived by many stakeholders as highly problematic: for example, evaluation mainly based on bibliometrics is perceived as often marginalizing research with potential societal contributions – thus clashing with policies fostering societal missions or challenges.

So deeply have these instruments become entrenched in contemporary governance, that it is difficult realistically to envisage a drastic reduction in the short term. This is first, because existing indicators have become "naturalized," fitting with the mainstream view among most researchers that "research quality" should be assessed on "internal" scientific criteria rather than discussing the relative value of broader contributions (Weinberg, 1962). Second, because NPM has become so strongly embedded in many governance settings (in evaluation machineries, infrastructure, and experts' networks) that strong dependencies have developed in the mobilization of quantitative evidence. Third, because evaluations based on less rigidly quantified expert judgments are not necessarily less problematic. Experts also have particular understandings and interests (i.e. biases), and those groups with resources or social capital tend to be relatively over-represented in panels – as shown, for example in gender, linguistic, racial geographical, and class biases. Moreover, experts are likely to be influenced by informal use of mainstream indicators, even if evaluations are formally based only on qualitative expertise (Kelly & Burrows, 2012).

Therefore, rather than advocating avoidance of indicators, most recent reform movements propose "responsible uses of quantification": where various forms of indicators and modeling are employed to support, but not substitute, expert judgment (Hicks, Wouters, Waltman, de Rijcke, & Ràfols, 2015; Saltelli et al., 2020; Wilsdon et al., 2017). The strategy we suggest is along the lines of statactivism (Bruno, Didier, & Vitale, 2014): counteracting traditional and conventional scalar indicators with new forms of quantification that illuminate the inconsistencies of narrow "performance indicators" and offer more plural alternatives. Thus, our proposal subverts the view of indicators from tools of control to tools of emancipation – thinking of indicator frameworks as Trojan horses that can be planted in evaluation processes for opening up critical debates and perspectives (Stirling, 2016).

This chapter explores how quantification can be developed and embedded in evaluation so that it offers "plural and conditional" perspectives, both to the evaluator and associated wider debate. Such practice is "plural" because quantification accommodates multiple perspectives in symmetrical ways. It is "conditional" because the resulting numbers are not presented as definitively unqualified, but as inextricably dependent on their contexts. An evaluator is still free to assess according to whichever perspectives she (or other relevant actors) can justify as being more appropriate. But this necessity for justification – and stimulus of wider critical scrutiny – adds crucial additional dimensions of rigor and accountability.

In suggesting this "plural and conditional" approach, we follow Andy Stirling's wider advocacy of practices for "opening up" social appraisal (the means by which society at large comes to apprehend alternative possible choices). This contends that both quantitative and qualitative approaches to evaluation (and appraisal more generally) can be used either for opening up or closing down debate (Stirling, 2008). Both functions are important, each is unavoidable, and (depending on context and perspective) either can have value, but a particular emphasis is warranted on "opening up" – under arguably any view – because powerful interests and dynamics of justification tend to introduce such a strong bias towards "closing down" (Stirling, 2012). Balancing these pressures therefore becomes a matter of rigor.

Accordingly, we argue that although indicators are currently mostly used to close down debate and endorse assessments shaped by dominant framings, alternative usages of quantification can also help foster higher quality public debates, make injustice more visible and enabling recognition for undervalued activities (Bruno et al., 2014; Lehtonen, Sébastien, & Bauler, 2016; Rottenburg & Merry, 2015, p. 25).

In particular, we propose that more "plural and conditional" indicators can help in making visible that notions of performance are intrinsically and fundamentally conditional on the particular perspectives or assumptions through which they are framed. It is therefore not just the resulting numbers that are important, but also a qualitative appreciation for the values and interests embedded within them (Pielke, 2007; Stirling, 2010).

We will focus here specifically on the context of research evaluation, although associated arguments apply across a diversity of sectors and evaluative conditions in social appraisal. The discussion is particularly relevant in situations in which scalar indicators have been applied to complex issues. This narrowing of vision has happened in instances of evaluation in sectors as disparate as

research, environment, education, or health: for example, scientific forestry in eighteenth-century Prussia (Scott, 1998), police statistics in New York (Bruno et al., 2014, p. 205), clinical practice in New Labour Britain (O'Mahony, 2017), or public health in the Global South (Adams, 2016).

The uses of indicators in research evaluation

Research assessment is an arena in which growing practice of "governance by indicators" has been especially diverse and intense (Burrows, 2012). From university rankings (van Raan, 2005), to performance-based funding systems (Hicks, 2012), or to individual-level assessments (Wouters, Glänzel, Gläser, & Ràfols, 2013), the use of "metrics" has become pervasive in research evaluation.

Thus, the professional incentive structure for researchers generally relies on a dominant framing according to which research performance may satisfactorily be characterized almost exclusively in relation to metrics of international publications. In these terms, "productivity" is associated simply with the number of publications churned out per researcher and research "quality" is associated merely with the number of citations per paper.[1] As bibliometric indicators became progressively established as a social institution and as infrastructure, the most popular indicators, such as journal impact factor (JIF) and the Hirsch (h-) index, became "naturalized as instantiations of quality irrespective of the methodological critiques by professional scientometricians" (Wouters, 2014, p. 58).

By implicitly insisting that research quality can be "measured" in the same way all around the world, these universalistic notions essentially assume that all research has the same purpose. Yet research managers and evaluators have known for decades that research "quality" is understood differently across contrasting scientific communities and depends on the contextual goals of the research (Roessner, 2000; Weinberg, 1962). Different notions of value apply, for instance, to research variously aiming to solve problems around local stakeholders' living conditions; provide policy advice on highly politicized social issues; foster public debates in uncertain areas of technology policy; enhance understanding of divergent priorities and interests in fields like education; or address narrow canonical disciplinary puzzles within academic settings (Chavarro, Tang, & Ràfols, 2017; Dahler-Larsen, 2019, p. 129). Notions of "quality" may have as many meanings in research as in other areas of culture (Dahler-Larsen, 2019, p. 4; Heuts & Mol, 2013).

It is for these reasons that uncontextualized uses of science and technology (S&T) indicators have been widely criticized (Feller, 2002, 2013; Weingart, 2005). Many reform initiatives have been launched, including the *San Francisco Declaration on Research Assessment* (DORA, 2013), the *Leiden Manifesto* (Hicks et al., 2015), and *The Metric Tide* (Wilsdon et al., 2017). As a result, research assessment is an area where issues around pluralization in the use of quantification has already been widely discussed (Barré, 2010, 2019; Lepori, Barré, & Filliatreau, 2008; Ràfols, 2019), with a number of prominent experiments being carried out (Benedictus, Miedema, & Ferguson, 2016; Lebel & McLean, 2018).

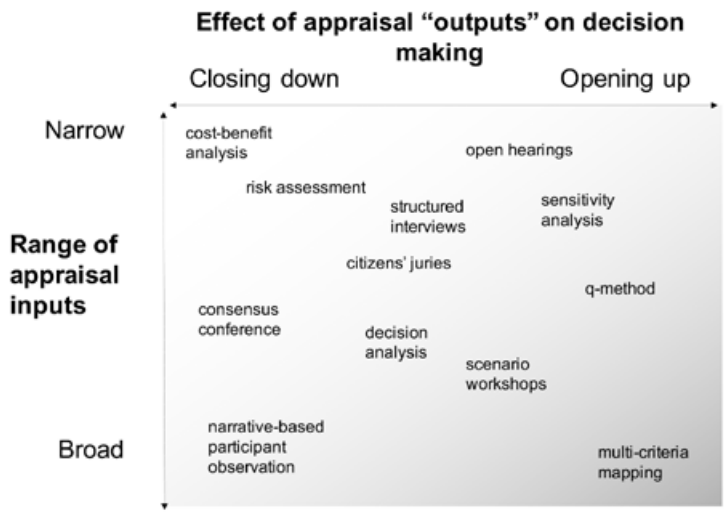
The particular way in which most efforts have sought to improve the robustness of measurements has lain in broadening out the range of inputs used in evaluations. In pursuing this, analysts have reverted to an early insight (subsequently neglected in much indicator activity) that assessments should rely on multiple sources of data that may provide "converging partial indicators" (Martin & Irvine, 1983). The broadening of inputs is facilitated by an avalanche of technical developments. First, the possibility of using different data sources stemming from the multiplication of traces now left over the cyberspace – including new publications databases such as *Microsoft Academic* or *Dimensions* (Visser, van Eck, & Waltman, 2020), and databases such as *Altmetrics.com* on uses or mentions of publications in social media or policy documents (Wouters, Zahedi, & Costas, 2019b). Second, many new tools have emerged for data visualization (e.g. Hans Rosling's *Gapminder*, commercial Tableau, or open source R statistics visualizations[2]), in particular for mapping large networks such as Gephi or VOSviewer[3] (Börner, 2010).

While this "broadening out" of the range of data used as "inputs" in evaluation is commendable, we suggest that a second – complementary and independent – dimension should also be considered. This focuses not on inputs to appraisal (like research evaluation), but on the "outputs" to evaluators and wider policy debates – attending to the extent to which these "open up" appreciation for contrasting conceptualizations of the phenomena under scrutiny. It is here that more "plural and conditional" communication of indicators can allow evaluators for more rigorous attention to alternative strategic considerations (Leach, Scoones, & Stirling, 2010; Stirling et al., 2007).

Opening up versus closing down in appraisal processes

Let us define appraisal as "the ensemble of processes through which knowledges are gathered and produced in order to inform decision-making and wider institutional commitments" (Leach et al., 2010). In our case, these appraisal processes are carried out through tools, methodologies, and approaches – quantitative or qualitative, analytical or participatory – that inform, and thus strongly shape, the outcomes of evaluations.

We can distinguish two dimensions in any appraisal process, as illustrated in Figure 10.1 (Stirling et al., 2007). The first dimension, the "range of appraisal inputs," refers to the scope, extent, and depth with which appraisal includes few or many different types of knowledge to describe the phenomena under scrutiny. The second dimension, the "effect of appraisal outputs in decision making," refers to the degree to which the outputs of appraisal facilitate "closing down" debate or, on the contrary, provide plural interpretations of the phenomena and thus foster "opening up" deliberation between contrasting options.



Note: Conventional uses of methods tend to fall in certain areas.
Source: Stirling et al. (2007, p. 57).

Figure 10.1 Schematic representation of the breadth of inputs in appraisal and the effect of outputs on decision making

Although typically highly diverse in their potentialities, distinct cultures of practice serve to lead different methods to occupy distinct spaces in Figure 10.1. Some methods build on smaller or larger ranges of inputs. Some techniques facilitate "closing down" appraisal by establishing an absolute ranking of "best" choices, while others foster "opening up" by allowing evaluators to compare and contrast how different assumptions in analysis may result in divergent rankings of options.

Along the vertical axis, one may argue (*ceteris paribus*) that appraisals which "broaden out" the range of inputs will tend on balance to be more comprehensive and thus more robust. If resources allow, efforts to increase breadth are thus generally desirable. This expansion of the range of inputs is particularly important when working with indicators, given that "the increasing importance of quantitative evidence leads to a situation in which only those operations which are counted and can be counted, count at all, and that qualitative and more complex operations will receive less and less attention" (Rottenburg & Merry, 2015, p. 20).

Along the horizontal axis, there is less *a priori* basis for normative preference. Policy processes typically yield contrasting moments in particular settings for "opening up" or "closing down" debates during an evaluation. These may of course be viewed differently, with important implications for the broad families of techniques that might legitimately be preferred in any given context. Yet (as discussed in the introduction), these specific routine dynamics in particular areas take place against a wider backdrop in which deeper and broader pressures for decision justification lead to a general bias towards methods for closing down (Stirling, 2019).

It is not only the case, therefore, that deliberate attention must be given to "opening up" the issues and perspectives in question prior to policy closure within any particular setting. At least equally important in the interests of rigor and accountability are that strenuous efforts must be made – and institutionalized – in order to balance the bias imposed by policy incumbents towards closure. It is on this basis that one may argue also from a standpoint of rigor, that where indicators have been used expediently to circumvent open scrutiny or democratic agency, a particular premium emerges across diverse political perspectives, for a premium on opening up (Dahler-Larsen, 2019, pp. 217–18).

Despite countervailing technical potentials, the cultures around methods like cost-benefit analysis tend to lead these to consider fewer relevant issues and provide ranked outputs that highlight the preferable choices thus facilitating a closing down of discussion across options. Thus, cost-benefit analyses

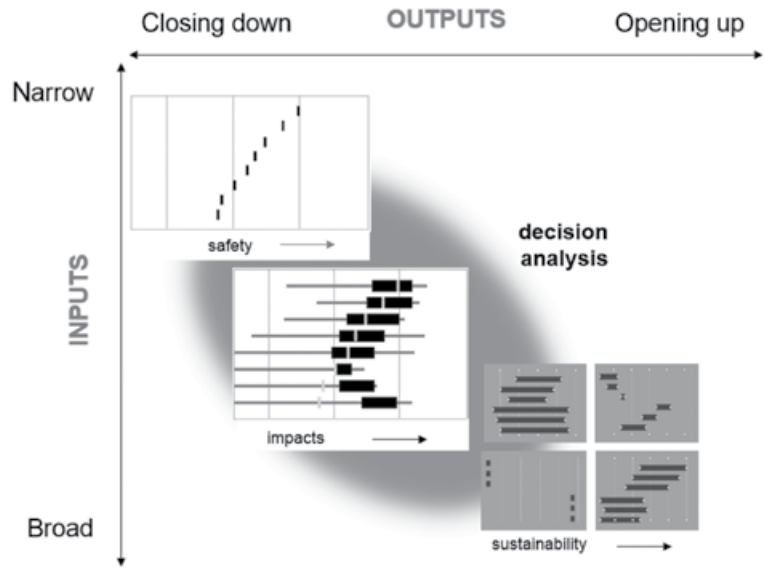
(upper-left of Figure 10.1) are often used to justify infrastructural decisions such as dams, by making some issues such as economic costs and benefits visible, while neglecting aspects not easily amenable to quantifications, such as effects on gender or cultural identities (Leach et al., 2010).

Methods such as open hearings or unstructured interviews (upper-right of Figure 10.1) may rely on small samples of views (thus are narrow), but they may have an opening-up effect if they introduce a diversity of perspectives. On the other end, consensus conferences (lower-left of Figure 10.1) may provide a variety of disparate views on an issue, but by definition, the focus on "consensus" means that the output is likely to facilitate making a decision, rather than further debate. However, the position of methods in the space of "range of inputs versus effect on outputs" depends on the specific use that is made.

One particular way of opening up decision making is to question the object of appraisal – i.e. what is to be evaluated. For example, should the evaluation consider research quality according to the immediate outputs (with indicators such as publications), intermediate outcomes (e.g. with indicators related to use by stakeholders), or the societal impacts (with indicators such as estimated contribution to health/wellbeing)? In methodology, *Research Quality Plus* efforts were made to make use of multiple understandings of the object of appraisal in order to judge the quality of development research (Lebel & McLean, 2018).[4] Rather than assuming the mainstream indicator, it would be worth having an explicit discussion on choice or keep a multidimensional description. Another issue concerns the units of analysis. For example, quantitative clustering may question existing classifications: clustering of researchers might show clusters very different from the institutional groups suggested following bureaucratic guidelines.

Further, one should consider the very different ways of using the same methods. Figure 10.2 provides the example of "decision analysis" to illustrate how different designs and implementations of a method can alter the breadth and openness of the method and thus change its position in the scheme shown in Figure 10.1. For example, decision analysis may focus on human safety as the only criterion to be considered, on the basis of scores provided by experts without uncertainty range. In this case, the options can be clearly ranked, and the method can be located in the upper-left side of the graph (Stirling, 2015, p. 26). Yet, a decision analysis process can also be implemented taking a wider range of impacts into consideration (human safety, environmental impact, cultural impact on populations affected) and include the uncertainty ranges, given the difficulties of estimating these impacts. As a result, as shown in the middle of Figure 10.2, now the options are not clearly ranked. There is thus

first a broadening out in inputs as impacts beyond human safety have been considered, and second, an opening up of outputs because by making uncertainties and ambiguities explicit there is no clear preferred option and policy debates become more relevant.



Note: This illustrates the conditional position of methods in the graph, although conventional uses of methods tend to fall in certain areas as shown in Figure 10.1.
Source: Stirling et al. (2015, p. 26).

Figure 10.2 Relative position of decision analysis in terms of breadth of inputs versus effect of outputs, under different uses of this method

Finally, the lower-right of Figure 10.2 gives the example of Multicriteria Mapping (MCM) (Coburn & Stirling, 2016).[5] This is a sophisticated hybrid quantitative-qualitative version of decision analysis which, instead of aggregating participants' views on the pros and cons of a range of options, uses a range of graphical and narrative tools to help "open up" greater appreciation of the reasons for what will typically remain persistent irreducible uncertainties and ambiguities. At every stage, MCM prioritizes the agency of participants themselves to frame issues and define the scope of appraisal in whichever ways they judge to be appropriate – thus broadening out appraisal to flexibly accommodate a full range of salient "inputs." Crucially, however, MCM also prioritizes various means to visualize each perspective separately, and so explore specific

reasons for differences. For instance, a comparison between charts shows the divergent perspectives' impacts and their uncertainties – thus highlighting different values by experts yielding different assessments. This helps to enable the "opening up" of the outputs of appraisal, equally for decision makers and to wider policy debates.

Whether facilitated by MCM or some other method of this kind, it is this kind of approach that is required in order to realize the quality of "plural and conditional" appraisal discussed in the introduction. The results obtained are explicitly "plural" both because each perspective is encouraged to highlight its own uncertainties concerning option ordering (rather than aggregate a single preference) and because contrasting such ordering of options is clearly associated with divergent real-world perspectives, each meaningful in different ways to the policy debate in question. And these results are rigorously "conditional," because each ordering is clearly associated with the subjective conditions which give it meaning, with rich qualitative information in this regard available to deepen and qualify the quantitative picture.

For the purpose of our present discussion on the use of indicators in evaluation, it is important to observe that, not only can similar methods occupy different positions depending on how they are implemented, but also that expert-analytic and participatory-deliberative methods are quite evenly distributed over Figure 10.1. This reflects longstanding appreciation that both analytic (often quantitative) and participatory (often qualitative) methods can – equally and in different ways – each be used alternatively to close down or to open up the policy processes that they inform.

This observation offers an important corollary to longstanding historical evidence for the ways in which analytical methods in particular (which tend to be quantitative) are often used to shut down debate and justify decisions (Porter, 1995; Rottenburg & Merry, 2015). The clarity and prominence of this evidence can sometimes lead to the assumption that more participatory or qualitative approaches are somehow intrinsically more suited to opening up, whilst analytical and quantitative approaches are inexorably all about closing down. To be fair to quantitative analysis, however, it should be pointed out that this is not necessarily the case (Stirling, 2008).

Indeed, it may be that it is more the epistemic authority of a quantitative-analytic idiom in contemporary policy cultures that makes these techniques preferable for interests wishing to justify closure. In cases where more qualitative and deliberative methods are used in policy making, pressures for closure are typically barely less evident – as reflected in the emphasis on particular interpre-

tations of analysts, or on ostensibly prescriptive "verdicts" and "consensus" in much participatory practice. Of course, these qualitative-deliberative methods can be used to illuminate contrasting interpretations or perspectives. But so too can analytic-quantitative approaches be used to map out the plural implications of diverse assumptions or framings. Experiences such as those narrated by the *Statactivism movement* (Bruno et al., 2014) show how alternative forms of quantifications can challenge incumbent perspectives and open debate.

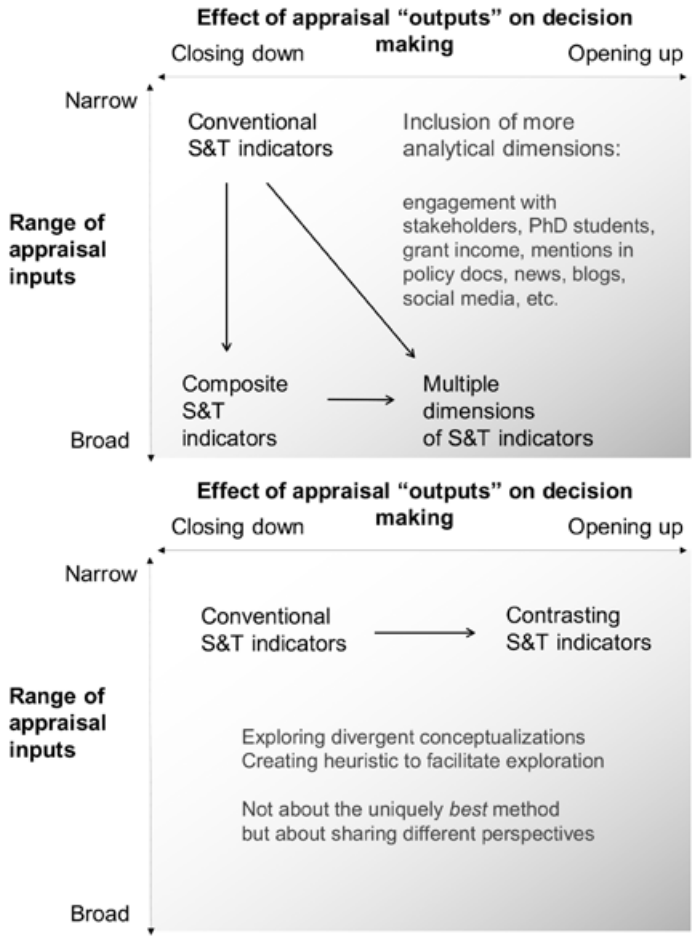
In summary, judgments over whether methods do, or ought to, open up or close down or not depend strongly on the design of the appraisal, its context, and the perspective under which these are viewed. As relevant to our present focus on quantitative evaluation as to other areas of appraisal, Stirling and colleagues proposed the notion of "empowering designs" for methods that aim at eliciting and foregrounding perspectives that are otherwise relatively marginalized. They contend that "inclusion" should go beyond the participation of excluded groups and extend to a symmetrical analytical treatment of alternative perspectives, thus facilitating processes of negotiation between actors on the values and the politics of appraisal (Leach et al., 2010).

In the following section, we will present some examples to illustrate how quantitative approaches can be used for opening up in research assessment.

"Broadening out" and "opening up" research evaluation with science and technology indicators

Where are indicators in research evaluation "positioned" in terms of the schematic representation of breadth of inputs versus openness of outputs introduced in Figure 10.1? We contend that they often lie in the upper-left corner, perhaps slowly moving toward the center-left, as illustrated in Figure 10.3. Conventional indicators in research evaluation are generally based in few inputs (mainly publications) and they are generally used as information to facilitate expediency in decisions, i.e. to close down notions of performance and associated debates.

However, following the discussion in the previous section, we will argue that S&T indicators can play a role in fostering pluralism rather than closing down



Note: Top: inclusion of more indicators, covering different analytical dimensions, leads to "broadening out" of evaluation. This leads to "opening up" when these different dimensions are shown explicitly. However, there is no significant "opening up" when this is followed by aggregative techniques such as composite indicators. Bottom: another route to "opening up" is to create contrasting indicators of the same analytical dimension under consideration, e.g. contrasting notions of bibliometric performance, of which convergence or divergence of insights can be discussed.

Figure 10.3 Illustration of types of shifts towards more plural use of science and technology indicators

perspectives. Three types of shifts can support more emancipatory use of S&T indicators:

1. Inclusion of more analytical dimensions (broadening out) while avoiding the use of aggregative techniques such as (simplistic) composite indicators (Figure 10.3 top).
2. Development of contrasting indicators (opening up) for analyzing the same issue, thus facilitating reflection on appropriate framing and analytical choices (Figure 10.3 bottom).
3. Shift to participatory dynamics (from indicators to indicating) so that quantification is contextualized in the goals, locations, and values of the specific evaluation.

Broadening out by including more analytical dimensions in indicators

As discussed in the second section, the problematic use of indicators in research evaluation has led to a backlash in the use of the more simplistic indicators such as JIF. This reaction against conventional indicators prompted a search for indicators that would capture the blind spots of scientometric measures, such as indicators of social contributions (or *impact*) of research (Molas-Gallart, Salter, Patel, Scott, & Duran, 2002) or indicators of Open Science (Wouters et al., 2019a). A parallel boom in the use of electronic platforms has led to a large expansion of data available for assessing research activities, in particular proliferation of indicators capturing non-conventional aspects of researcher performance (Pontille & Torny, 2013).

However, the availability of data for broadening out does necessarily translate into a pluralization of research evaluation – for example, when the closing down of conventional bibliometric indicators is substituted by the closing down of new *Altmetric* or *Open Science* indicators that follow the same integrative productivist logic (Robinson-García, van Leeuwen, & Ràfols, 2018).

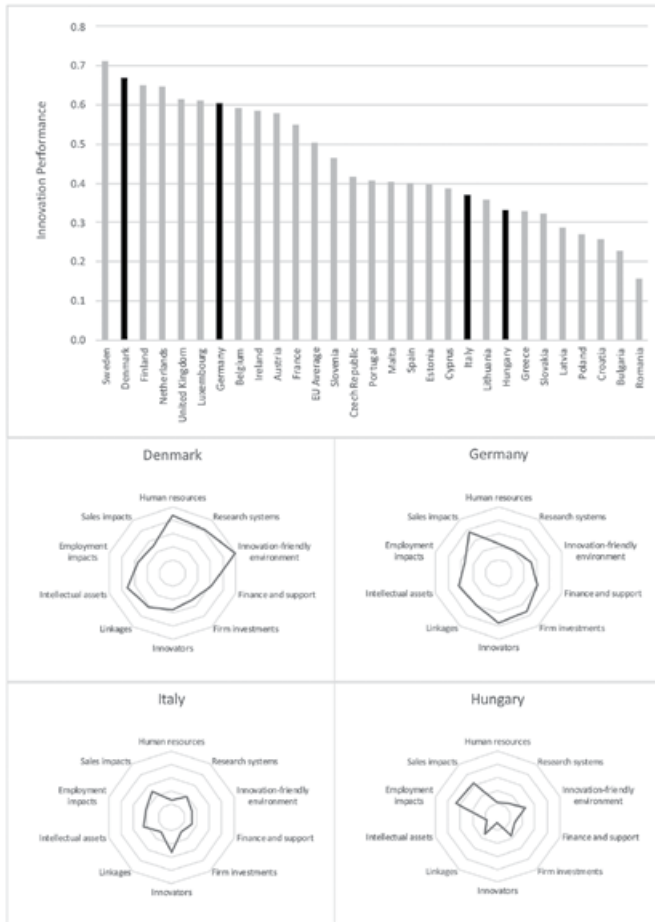
University ranking providers exemplify how analyses considering various dimensions do not necessarily lead to more pluralistic understandings. Let us leave aside for a moment that the data and the methodologies behind these rankings are, to put it mildly, rather problematic (van Raan, 2005). Most rankings are based on very distinct analytical dimensions, such as the quality of education, of research, the international outlook, or the industry income. Yet in the end, much of the benefit for improved understanding that might arise from this broadening out of consideration across more different dimensions is then lost when all these dimensions are folded into a single composite index.

With contrasting equally reasonable protocols for aggregation typically yielding radically different index orderings, particular chosen parameter structures will at best be arbitrary and at worst vulnerable to gaming or capture.

This closing down in spite of richer information also occurs with more rigorous analysis like the *European Innovation Scoreboard*. In 2017, this consisted of 10 analytical dimensions based on 27 indicators (between two and three indicators per dimension) (Hollanders & Es-Sadki, 2017). These 27 indicators were summarized in a single scalar score, effectively "closing down" debates on performance by univocally emphasizing a particular country as "most innovative." Such composite indices have been shown to be potentially misleading as "the scope for manipulation of scoreboards by selection, weighing and aggregation is great" (Grupp & Mogee, 2004, p. 1382; Grupp & Schubert, 2010). Yet, as shown in Figure 10.4, simply displaying the analysis in radar charts, rather than in one dimension, allows appreciation for the ways in which ostensibly similar aggregate scores may obscure very different profiles (compare, for example of Denmark versus Germany, or Italy versus Hungary).

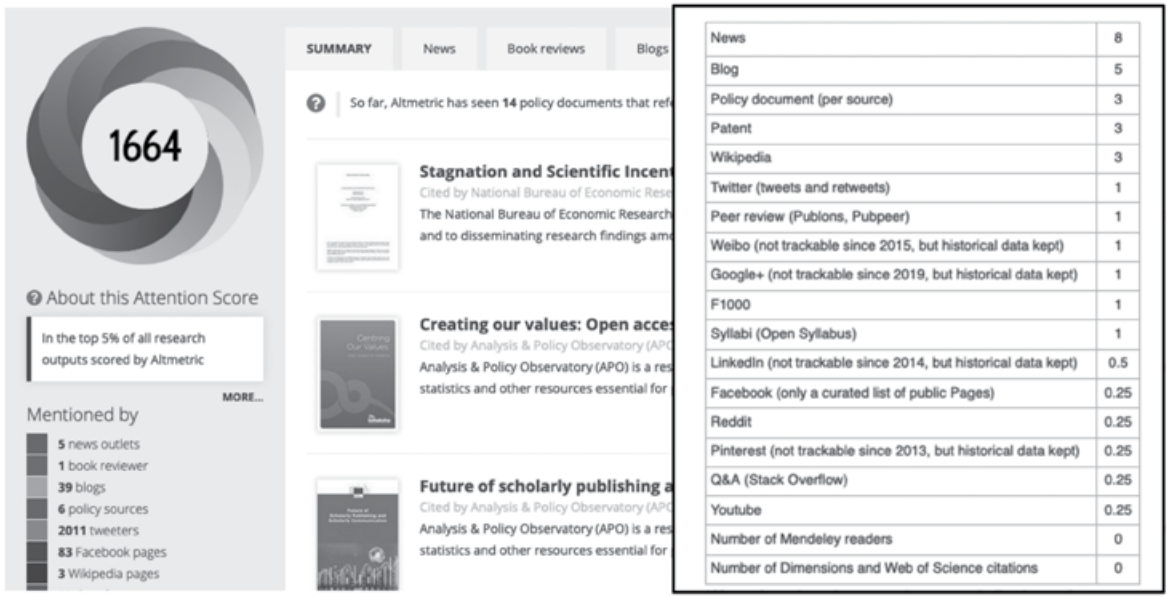
For our focus on research assessment, the development of *Altmetrics* indicators is paradigmatic to cast light on the challenges of broadening out given the political economy of research assessment. The initial proponents of *Altmetrics* were genuinely eager to pluralize research assessment with new "metrics" that could report activities invisible in conventional approaches, such as blogging and data or code sharing (Priem, 2014; Priem, Taraborelli, Groth, & Neylon, 2010). Indeed, in the last decade there has been a blossoming of scientific traces in the cyberspace: repositories of data, preprints and postprints, codes, databases analyzing mentions of academic work in social media, etc. One might thus have expected that the analysis of these traces would lead to consolidation of new indicators of social attention.

As a matter of fact, *Altmetric.com*[6] has been successfully marketing data on the attention generated in societal media by a publication as well as an indicator, the *Altmetric Attention Score*, which is a composite index giving different weights to mentions in news, blogs, policy documents, patents, *Twitter*, *Facebook*, and *YouTube* (as shown in Figure 10.5). These metrics are provided by *Altmetric.com* without standards for comparison, they have a very irregular coverage, and the meaning of their aggregate score is unclear. Therefore, they have not shown so far to be meaningful or reliable quantitative indicators for evaluation purposes (Robinson-García, Costas, Isett, Melkers, & Hicks, 2017; Sugimoto, Work, Larivière, & Haustein, 2017; Wilsdon et al., 2017).



Note: The European Innovation Scoreboard is a composite index of innovation "performance" that aggregates multiple analytical dimensions. However, its aggregate nature does not allow to see the different strengths by country. A simple radar chart makes explicit the contrasting profiles even for countries with a similar aggregate performance, as shown comparing Denmark against Germany, or Italy against Hungary.
Source: Ràfols (2019) based on Grupp and Schubert (2010). Data source: European Innovation Scoreboard (in Hollanders & Es-Sadki, 2017).

Figure 10.4 Visualizing multiple dimensions in radar charts



Note: Left: example of the Altmetric Attention Score of an article by *Altmetric.com*. The score is in the center of the doughnut. Below the doughnut, you can see the actual counts in different dimensions such as news, blogs, etc. By clicking on the tabs one can look up the actual news, blogs, or tweets in which the article was mentioned. Right: the weights used by each dimension of the Altmetric Attention Score. See details in https://dimensions.altmetric.com/details/3931894.

Figure 10.5 Example of the Altmetric Attention Score of an article by *Altmetric.com*, and the weights used by each dimension of the Altmetric Attention Score

This said, it should be acknowledged that the information provided by *Altmetric.com* is rich and can be very useful in tentatively exploring (by clicking tabs and digging into details) whether and why a publication has generated interest. Not only does this help illuminate the kinds of attention received (whether news, policy, or blogs), it also allows users to search specific instances. Thus, it is arguably mostly in non-aggregated forms (as indicators within different dimensions like news, blogs, policy documents, patents, etc.) that *Altmetric* data can best be used to pluralize understandings on the part of research policy audiences (i.e. in the diagonal shift to multiple dimensions in Figure 10.3).

In this sense, we view the Altmetric Attention Score as an example of one of the main challenges in broadening out. It is precisely when a wider range of inputs are included that pressures from managerial interests to perform a "hard" composite index can lead evaluative practices unduly to ignore problems of incommensurability in the component indicators and sensitivity of results to aggregation protocols. Thus, in spite of undoubted good will (Priem et al., 2010), we view *Altmetrics* as an interesting development for exploratory analysis (Costas, Honk, Calero-Medina, & Zahedi, 2017; Noyons, 2019), but with very questionable impact so far in research assessment as a result of a decontextualized implementation, in the same accountability (and "bean counting") tradition of conventional bibliometrics (Barré, 2019; Ràfols, 2019).

Opening up by considering contrasting indicators of the same property

Let us now turn from practices of broadening out S&T indicator inputs (as shown in the bottom of Figure 10.3) to challenges of opening up S&T evaluation without necessarily adding very large arrays of data sources. How can it be possible to foster more plural analyses, even when attention is dominated by particular sources – such as a specific bibliometric database? How can quantitative studies capture and convey diverse perspectives on a given issue, even by reference to the same body of data?

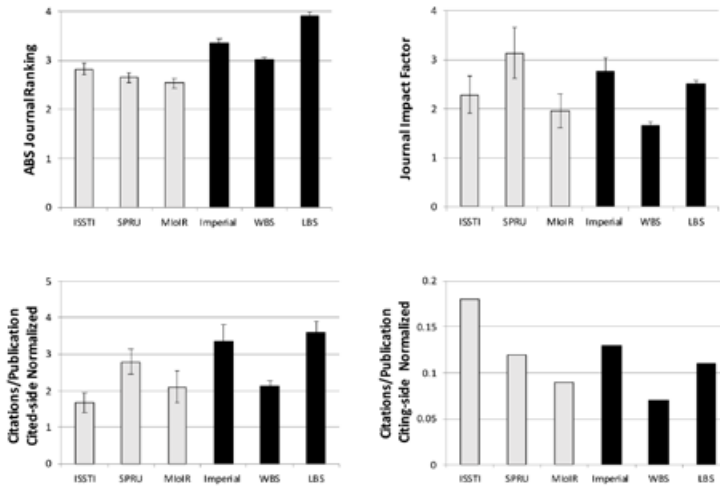
It is crucial to the distinction made here between (related but often effectively quite independent) qualities of "broadening out" and "opening up," to appreciate that "opening up" can be undertaken without necessarily "broadening out." All that is required is an openness to exploring contrasting operationalizations of some single property of interest. In other words, even with narrow inputs, tools can be developed that help evaluators scrutinize how different

conceptualizations and associated mathematical operationalizations may yield contrasting results with the same data. By investigating how different assumptions lead to different methods and rankings (even using only a single indicator and dataset), the analyst can provide "plural and conditional" advice.

By helping to cultivate a policy culture that is more generally reflective over the importance of uncertainty and variability and more reflexivity over the normative – ethical and political – aspects of apparently technical analytical choices, a particular exercise in opening up may even help to nurture a greater general attentiveness and responsiveness even to parameters that were not included in its own analysis. Both the practice of evaluation and associated policy debates may thereby be made more rigorous and accountable.

Let us take the core notion of "research quality" (which, it may be recalled, Martin and Irvine established in 1983 that bibliometrics could *not* address!). In conventional bibliometric analysis, research quality is interpreted as referring only to the academic perceptions of value. This is then operationalized using publication data, but in diverse ways: in terms of journal rankings (a disciplinary list), in terms of JIF, and in terms of citations, which in turn can be normalized (made commensurable) according to field of the article receiving or giving the citation (cited-side normalized or citing-side normalized). These yield radically different understandings of "quality." A journal ranking produced by an academy (e.g. the UK Association of Business Schools) counts as quality publishing in the most prestigious journals of a discipline. The JIF assumes that quality is related to the citation impact of the journal. Cited-side normalization considers that citations rather than journals define quality and that all citations are equal. Citing-side normalization considers that attracting citations from fields that cite little is more valuable.

While these different conceptualizations and corresponding operationalizations can be easily understood as diverse, most people will be surprised when shown, as in Figure 10.6, that these different choices lead to strikingly different results. It is generally assumed that these choices may change the details, but the relative order of performance will remain stable. In this case, we compared the bibliometric performance of three interdisciplinary units of innovation studies with three schools of business and management (Ràfols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012b). Given that some of the units under analysis were highly interdisciplinary centers, the results were greatly affected by specific operationalizations. This lack of congruence has sometimes been studied regarding technical issues such as field normalization techniques (Adams, Gurney, & Jackson, 2008; Zitt, Ramanana-Rahary, & Bassecoulard, 2005) or dimensions such as language (van Leeuwen, Moed, Tijssen, Visser, &



Note: The acronyms represent different institutes and schools as reported in Rafols et al. (2012b); units of science and innovation studies (left, in grey) and business and management schools (right, in black).
Source: Ràfols et al. (2012b).

Figure 10.6 Contrasting results using different measures of research performance of university units of Science and Innovation Studies and Business and Management schools

van Raan, 2001), but it is seldom debated in evaluative applications – perhaps with the exception of the biases generated by disciplinary (Hicks, 1999) and geographical coverage of databases (Chavarro, 2017; Vessuri, Guédon, & Cetto, 2014).

For concepts such as interdisciplinarity, that enjoy a conspicuous lack of consensus, exploring contrasting indicators is even more important (Wang & Schneider, 2020). When results are convergent, agreement provides robust evidence of insights (Ràfols et al., 2012b). When results are divergent, interpretation is challenging and might be disconcerting against some expectations (Digital Science, 2016). However, it should not be assumed that divergence means that the contrasting indicators are necessarily invalid. Rather, it may be that different interpretations of interdisciplinarity provide different insights. In this case, the actors concerned with the evaluation need to engage with their particular understandings of "interdisciplinarity" so as to choose the specific processes of operationalization that they find relevant for their case and context. In opening up the operationalization of the indicators as a plural

and conditional process, we achieve the key step of moving from indicators to indicating (Marres & de Rijcke, 2020).

These bibliometric examples on operationalizations of "research quality" illustrate that our proposal for "opening up" should not be seen as an impractical call for ever more inputs and ever more outputs. One should not interpret that "opening up" is about giving more indicators – it is about adding the minimum number of indicators that will force decision makers to consider the relevant evaluative options rather than thoughtlessly grab the easiest naturalized indicator (let's say the JIF). Far from requiring postponement of decisions, opening up can help avoid cost and delay caused by protracted controversies provoked by unreasonable attempts to assert single indicators that fail to reflect from the outset a requisite range of salient issues.

Where there is only one indicator (and associated implicit framing) of a property that is widely recognized as contentious, it will often be enough to add just a second indicator that provides a contrasting perspective in order to avoid reflection on which of the two framings and associated indicators are more appropriate in a given evaluative context. Let us be clear that we believe that parsimony is very important in policy indicators in order to allow transparency. But apparent simplicity should not be achieved through suppression or conflation of relevant evaluation dimensions.

From indicators to indicating: engagement for plural and conditional advice

To counter the use of indicators as rigid tools that capture only narrow understandings of the issues evaluated and then marginalize certain options, we have proposed to build on Stirling's framework of "empowering designs" (Stirling et al., 2007): this is to develop and apply quantification in ways that (1) broaden out the scope of knowledge gathered, and (2) have a pluralizing effect (i.e. open up) in the evaluative process. Broadening out involves considering more analytical dimensions – opening up consists in actively fostering more critical debate, rather than closing it down. Each can occur quite independently, with a useful degree of opening up being possible even without a corresponding broadening out. We have also argued, on the other hand, that broadening out without opening up (e.g. in university rankings on *Altmetrics*) does not result in a significant pluralization of evaluation.

By taking more analytical dimensions into account, and/or by exploring contrasting perspectives on these dimensions, we are effectively expanding the potential insights gained from indicators (or indicating) in the evaluation. How can the evaluator come to make decisions under these more plural circumstances? How can evaluation proceed without a clear set of indicators?

Pielke (2007) argued that under conditions of uncertainty and lack of consensus, it is not possible in scientific advice to separate knowledge formation (in our case: the construction of indicators) from decision making (in our case: evaluation). Experts on the world of indicators, Rottenburg and Merry reached a similar insight: "it is impossible to separate the concrete processes of measuring [the construction of the indicators] from the actual use of the indicators [in decision making]" (2015, p. 30).

This means that the choice of ostensibly objective indicators for a given evaluation is inevitably related to underlying intrinsically subjective "valuations." This is an intuition with which we are all familiar when dealing with mundane objects such as tomatoes (Heuts & Mol, 2013) – different criteria of "quality" are applied depending on the expected usage and preferences on taste, texture, color, etc. In other words, the choice of indicators is conditional on the values, which is why expert analysts should offer "plural and conditional advice" – with multiple indicators and values under each explicitly conditional on assumptions appropriate under relevant values and contexts. Whilst the apparent parsimony of aggregate indices may superficially look like a virtue in scientific advice, this may conceal an intractable and volatile complexity of hidden contingencies (Grupp & Mogee, 2004). In this sense, a simple general heuristic of "opening up" may offer a rigorous more robust form of parsimony (Stirling, 2010).

In a previous publication, one of us argued that from the adoption of this plural and conditional framing, it follows that indicators have to be constructed with the participation of stakeholders in the middle of the social world, perhaps during evaluations – what we called "indicators in the wild" (Ràfols, 2019) after Callon's "research in the wild" (Callon, Lascoumes, & Barthe, 2001). Waltman and van Eck (2016) proposed "contextualized scientometrics" as a form of scientometrics that would allow "users" to shape the quantitative analysis with their contextual knowledge. Marres and de Rijcke have pointed out that this shift from off-the-shelf, universal indicators to tailored contextual indicators means that we move from a product (indicator) to a process (indicating).

> We describe this approach as *indicating* to highlight something that each of the four terms above [participatory, abductive, interactive, and designed] have in common:

> they frame the development and use of indicators as a process. This is key insofar as it enables us to understand the assembly of communities of interpretation as an on-going process, one that spreads out across the design and deployment of indicators. (Marres & de Rijcke, 2020, p. 1050)

It is during this process of "indicating" that closing down takes place. Thus, making efforts to open up indicators does not mean that decisions themselves remain open. Decisions can still be taken as needed. What is different is simply that the process of decision making is enriched by more reflective and explicit consideration to the rationales behind possible choices, without expedient closure of indicators allowing the obscuring of decision accountability. Collective deliberation might, in some cases, facilitate the construction of shared perspectives among stakeholders. But decisions will nonetheless still likely need to take place in the face of incommensurable perspectives and persistent stakeholder contention, each made explicit in contrasting preferences for indicators. In this light, opening up does not impede decision making, but merely enables decision makers to be more clear and transparent in explaining their choices. Even though some individual decision makers may sometimes prefer hiding behind indicators that claim to offer the "best" technocratic advice, wider public interests might hold that this kind of rigor and accountability should be routinely expected in mature democratic governance.

Of course, it does sometimes occur that decisions need to be made under conditions of uncertainty. For example, bibliometric indicators of individual researchers, while informative, are not reliable to fully assess fellowship candidates and reviewers do not agree on the ranking of candidates. Under these conditions it is advisable for assessment to recognize uncertainty and proceed with methods that embrace it (such as partially randomized selection[7]) rather than using indicators, which is likely to lead to systematic biases (e.g. favoring men, basic and fashionable topics) and indicator gaming (de Rijcke et al., 2016).

There are experiences of this shift towards participation of decision makers and stakeholders in the design and use of indicators. For example, the new evaluation framework developed by the Utrecht Medical Centre was created after a process that involved public debates (Benedictus et al., 2016). De Rijcke et al. (2019) advocate an approximation to evaluation ("evaluative inquiry") that espouses adopting the methods to the particular needs of an evaluation. These efforts towards participatory quantifications require the design of new methodologies that act as interfaces between different actors (Marres & Gerlitz, 2016). Marres and de Rijcke (2020) emphasize the value of building on

expertise in participatory methods, user studies, and design research in order to develop these methodologies.

International development is one of the policy areas that pioneered this participatory turn (Chambers, 1995), and where examples of practices of "participatory statistics" might be sought (see Holland, 2013, showing a variety of experiences). For example, the method Participatory Impact Pathway Analysis involves a variety of stakeholders in deciding, from the outset, what are the indicators that will be used during the monitoring and evaluation of an intervention (Douthwaite et al., 2007).

Conclusions

The use of indicators in evaluation (as well as in other social spheres) has become both pervasive and problematic. Conventional indicators facilitate closing down debate in evaluative processes by valuing an activity according to dominant analytical perspectives, for example publication productivity and citation impact in research evaluation. Indicators thus play performative roles, incentivizing and "guiding" both evaluands and evaluators towards particular understandings of "good" performance that tend to align with power.

In this chapter we have argued that while it is indeed the case that conventional quantification using scalar indicators has this blinkering effect, indicators can also be used to help support more plural evaluation and foster more productively critical debate. To achieve this shift towards indicators that foster perspectival diversity, we urge greater attention to two dimensions of design in the process of indicating. The first dimension, "broadening out," concerns the range of "inputs" taken into account in evaluation. The second, "opening up," relates to the "outputs" of quantifications, encouraging methodologies that enable attention to plural perspectives.

We have illustrated that even analytical tools as narrow as scientometric indicators leave room for evaluative usage that is more explicit about the dependence of analytic outputs on normative assumptions. We have shown that this "opening up" is distinct (and complementary) to the "broadening out" of the range of data inputs. We suggest that this move towards more situated and participatory use of quantitative evidence in evaluation implies a shift away from notionally universal *indicators* (as products) to more contextualized *indicating* (as process).

If conventional scalar indicators hold the "capacity to produce constitutive effects in such a way that conventional forms of democratic control are circumvented" (Dahler-Larsen, 2019, p. 218), the designs of quantification proposed here aim to illuminate instead a more democratic diversity of perspectives. We hope that these empowering designs can be creatively weaved into new policy contexts that allow quantification to challenge scalar instrumentalism and instead help foster democratic pluralism and accountability in evaluation.

Acknowledgments

This chapter builds on previous work and discussions with colleagues in the Science Policy Research Unit at the University of Sussex, Centre for Science and Technology Studies at Leiden University and at *Ingenio* (CSIC-UPV), Universitat Politècnica de València. An earlier, much shorter version of this manuscript was first published in the Proceedings of the 2012 S&T Indicators Conference (Ràfols, Ciarli, Van Zwanenberg, & Stirling, 2012a) and translated into Portuguese in the Proceedings of the 2017 Brazilian Meeting of Bibliometrics and Scientometrics.

Notes

1. It is worth remembering that the first (and then very controversial) studies using bibliometric indicators in research assessments spelled out in the abstract that analysis had to be done at the group level (rather than individual), that citations showed impact (rather than quality), that comparisons could only be made between "matched" groups, and that indicators were "partial" and only reliable when multiple indicators "converged" (Martin & Irvine, 1983). Notice, though, that the emphasis nonetheless lay in producing a "convergent" measure of a focal notion (like "scientific impact"), rather than in illuminating contrasting perspectives or measures.
2. See www.gapminder.org; www.tableau.com; http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html.
3. See www.vosviewer.com and https://gephi.org.
4. We thank one reviewer for suggesting this example. See also: www.idrc.ca/en/research-in-action/research-quality-plus.
5. See dedicated website at www.multicriteriamapping.com.
6. *Altmetric.com* is owned by Digital Science, a company of Holtzbrink group, which is also the owner of Springer-Nature.
7. See initiatives by the Volkswagen Foundation at www.volkswagenstiftung.de/en/funding/our-funding-portfolio-at-a-glance/experiment/partially-randomized-procedure.

References

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the Zoom: A Test of Zitt's Hypothesis. *Scientometrics*, 75 (1), pp. 81–95.

Adams, V. (ed.) (2016). *Metrics: What Counts in Global Health*. Durham, NC: Duke University Press.

Barré, R. (2010). Towards Socially Robust S&T Indicators: Indicators as Debatable Devices, Enabling Collective Learning. *Research Evaluation*, 19 (3), pp. 227–31.

Barré, R. (2019). Les indicateurs sont morts, vive les indicateurs! Towards a Political Economy of S&T Indicators: A Critical Overview of the Past 35 Years. *Research Evaluation*, 28 (1), pp. 2–6.

Benedictus, R., Miedema, F., & Ferguson, M. W. (2016). Fewer Numbers, Better Science. *Nature*, 538 (7626), pp. 453–5.

Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. Cambridge, MA: MIT Press.

Bruno, I., Didier, E., & Vitale, T. (2014). Statactivism: Forms of Action between Disclosure and Affirmation. *Open Journal of Sociopolitical Studies*, 7 (2), pp. 198–220.

Burrows, R. (2012). Living with the H-Index? Metric Assemblages in the Contemporary Academy. *Sociological Review*, 60 (2), pp. 355–72.

Callon, M., Lascoumes, P., & Barthe, Y. (2001). *Agir dans un monde incertain: essai sur la démocratie technique*. Paris: Seuil. English translation: Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an Uncertain World: An Essay on Technical Democracy (Inside Technology)*. Cambridge, MA: MIT Press.

Chambers, R. (1995). Poverty and Livelihoods: Whose Reality Counts? *Environment and Urbanization*, 7 (1), pp. 173–204.

Chavarro, D. (2017). *Universalism and Particularism: Explaining the Emergence and Growth of Regional Journal Indexing Systems* (PhD thesis). University of Sussex. Retrieved July 11, 2020 from: http://sro.sussex.ac.uk/id/eprint/66409/.

Chavarro, D., Tang, P., & Ràfols, I. (2017). Why Researchers Publish in Non-Mainstream Journals: Training, Knowledge Bridging, and Gap Filling. *Research Policy*, 46 (9), pp. 1666–80.

Coburn, J., & Stirling, A. (2016). *Multicriteria Mapping Manual – Version 2.0* (Manual). Science Policy Research Unit, Brighton. Retrieved July 11, 2020 from: http://sro.sussex.ac.uk/id/eprint/65615/

Costas, R., Honk, J. V., Calero-Medina, C., & Zahedi, Z. (2017). Exploring the Descriptive Power of Altmetrics: Case Study of Africa, USA and EU28 Countries (2012–2014). Presentation at the Science, Technology and Innovation Indicators Conference, September 6–8, Paris.

Dahler-Larsen, P. (2011). *The Evaluation Society*. Stanford, CA: Stanford University Press.

Dahler-Larsen, P. (2019). *Quality: From Plato to Performance*. London: Palgrave.

de Rijcke, S., Holtrop, T., Kaltenbrunner, W., Zuijderwijk, J., Beaulieu, A., Franssen, T., van Leeuwen, T., Mongeon, P., Tatum, C., Valkenburg, G., & Wouters, P. (2019). Evaluative Inquiry: Engaging Research Evaluation Analytically and Strategically. *Journal for Research and Technology Policy Evaluation*, 48, pp. 176–82.

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation Practices and Effects of Indicator Use: A Literature Review. *Research Evaluation*, 25 (2), pp. 161–9.

Desrosières, A. (2015). Retroaction: How Indicators Feed Back onto Quantified Acto. In Rottenburg, R., Merry, S. E., Park, S. J., & Mugler, J. (eds). *The World of Indicators: The Making of Governmental Knowledge through Quantification*. Cambridge: Cambridge University Press.

Digital Science (2016). Interdisciplinary Research: Methodologies for Identification and Assessment (Report). Retrieved November 10, 2020 from: www.mrc.ac.uk/documents/pdf/assessment-of-interdisciplinary-research/.

DORA (2013). San Francisco Declaration on Research Assessment. Retrieved July 10, 2020 from: https://sfdora.org/read/.

Douthwaite, B., Alvarez, S., Cook, S., Davies, R., George, P., Howell, J., Mackay, R., & Rubiano, J. (2007). Participatory Impact Pathways Analysis: A Practical Application of Program Theory in Research-for-Development. *Canadian Journal of Program Evaluation*, 22 (2), pp. 127–59.

Feller, I. (2002). Performance Measurement Redux. *American Journal of Evaluation*, 23 (4), pp. 435–52.

Feller, I. (2013). Performance Measures as Forms of Evidence for Science and Technology Policy Decisions. *Journal of Technology Transfer*, 38 (5), pp. 565–76.

Grupp, H., & Mogee, M. E. (2004). Indicators for National Science and Technology Policy: How Robust are Composite Indicators? *Research Policy*, 33 (9), pp. 1373–84.

Grupp, H., & Schubert, T. (2010). Review and New Evidence on Composite Innovation Indicators for Evaluating National Performance. *Research Policy*, 39 (1), pp. 67–78.

Heuts, F., & Mol, A. (2013). What Is a Good Tomato? A Case of Valuing in Practice. *Valuation Studies*, 1 (2), pp. 125–46.

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. (2012). Performance-Based University Research Funding Systems. *Research Policy*, 41 (2), pp. 251–61.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Ràfols, I. (2015). Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature*, 520 (7548), pp. 429–31.

Holland, J. (ed.) (2013). *Who Counts? The Power of Participatory Statistics*. Rugby: Practical Action Publishing.

Hollanders, H., & Es-Sadki, N. (2017). European Innovation Scoreboard 2017 (Report). European Commission, Brussels. Retrieved June 22, 2020 from: https://ec.europa.eu/docsroom/documents/24829

Kelly, A., & Burrows, R. (2012). Measuring the Value of Sociology? Some Notes on Performative Metricization in the Contemporary Academy. *Sociological Review*, 59, pp. 130–50.

Leach, M., Scoones, I., & Stirling, A. (2010). *Dynamic Sustainabilities. Technology, Environment, Social Justice*. London: Earthscan.

Lebel, J., & McLean, R. (2018). A Better Measure of Research from the Global South. *Nature*, 559, pp. 23–6.

Lehtonen, M., Sébastien, L., & Bauler, T. (2016). The Multiple Roles of Sustainability Indicators in Informational Governance: Between Intended Use and Unanticipated Influence. *Current Opinion in Environmental Sustainability*, 18, pp. 1–9.

Lepori, B., Barré, R., & Filliatreau, G. (2008). New Perspectives and Challenges for the Design and Production of S&T Indicators. *Research Evaluation*, 17 (1), pp. 33–44.

Marres, N., & de Rijcke, S. (2020). From Indicators to Indicating Interdisciplinarity: A Participatory Mapping Methodology for Research Communities in-the-Making. *Quantitative Science Studies*, pp. 1041–55.

Marres, N., & Gerlitz, C. (2016). Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *Sociological Review*, 64 (1), pp. 21–46.
Martin, B. R., & Irvine, J. (1983). Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio Astronomy. *Research Policy*, 12 (2), pp. 61–90.
Molas-Gallart, J., Salter, A., Patel, P., Scott, A., & Duran, X. (2002). Measuring Third Stream Activities (Report). Russell Group of Universities. Brighton: SPRU, University of Sussex. Retrieved July 11, 2020 from: http://ict-industry-reports.com.au/wp-content/uploads/sites/4/2013/10/2002-Measuring-University-3rd-Stream-Activities-UK-Russell-Report.pdf.
Noyons, E. (2019). Measuring Societal Impact is as Complex as ABC. *Journal of Data and Information Science*, 4 (3), pp. 6–21.
O'Mahony, S. (2017). Medicine and the McNamara Fallacy. *Journal of the Royal College of Physicians of Edinburgh*, 47 (3), pp. 281–7.
Pielke, Jr., R. A. (2007). *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge: Cambridge University Press.
Pontille, D., & Torny, D. (2013). La manufacture de l'évaluation scientifique. *Réseaux*, 1, pp. 23–61.
Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
Priem, J. (2014). Altmetrics. In Cronin, B., & Sugimoto, C. R. (eds) (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press.
Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved June 22, 2020 from: http://altmetrics.org/manifesto/.
Ràfols, I. (2019). S&T Indicators in the Wild: Contextualization and Participation for Responsible Metrics. *Research Evaluation*, 28 (1), pp. 7–22.
Ràfols, I., Ciarli, T., Van Zwanenberg, P., & Stirling, A. (2012a). Towards Indicators for Opening up S&T Policy. *STI Indicators Conference*. Available in Arxiv.
Ràfols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012b). How Journal Rankings Can Suppress Interdisciplinary Research: A Comparison between Innovation Studies and Business and Management. *Research Policy*, 41 (7), pp. 1262–82.
Robinson-García, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The Unbearable Emptiness of Tweeting about Journal Articles. *PLoS One*, 12 (8).
Robinson-García, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using Altmetrics for Contextualized Mapping of Societal Impact: From Hits to Networks. *Science and Public Policy*, 45 (6), pp. 815–26.
Roessner, D. (2000). Quantitative and Qualitative Methods and Measures in the Evaluation of Research. *Research Evaluation*, 9 (2), pp. 125–32.
Rottenburg, R., & Merry, S. E. (2015). *A World of Indicators: The Making of Governmental Knowledge through Quantification*. In Rottenburg, R., Merry, S. E., Park, S. J., & Mugler, J. (eds). *The World of Indicators: The Making of Governmental Knowledge through Quantification*. Cambridge: Cambridge University Press, pp. 1–33.
Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Pielke, Jr, R. et al. (2020). Five Ways to Ensure that Models Serve Society: A Manifesto. *Nature*, 582, pp. 482–4.
Scott, J. C. (1998). *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.

Stirling, A. (2008). "Opening Up" and "Closing Down": Power, Participation, and Pluralism in the Social Appraisal of Technology. *Science, Technology and Human Values*, 33 (2), pp. 262–94.
Stirling, A. (2010). Keep it Complex. *Nature*, 468 (7327), pp. 1029–31.
Stirling, A. (2012). Opening up the Politics of Knowledge and Power in Bioscience. *PLoS Biol*, 10 (1), e1001233.
Stirling, A. (2015). Developing "Nexus Capabilities": Towards Transdisciplinary Methodologies (Discussion paper). Science Policy Research Unit, Brighton. Retrieved July 11, 2020 from: http://sro.sussex.ac.uk/id/eprint/69094/.
Stirling A. (2016). Knowing Doing Governing: Realizing Heterodyne Democracies. In Voß, J. P. & Freeman, R. (eds). *Knowing Governance*. London: Palgrave Macmillan. https://doi.org/10.1057/9781137514509_12.
Stirling, A. (2019). How Deep Is Incumbency? A "Configuring Fields" Approach to Redistributing and Reorienting Power in Socio-Material Change. *Energy Research and Social Science*, 58, 101239.
Stirling, A., Leach, M., Mehta, L., Scoones, I., Smith, A., Stagl, S., & Thompson, J. (2007). Empowering Designs: Towards More Progressive Appraisal of Sustainability (Working paper 3). STEPS Centre, Institute of Development Studies, Brighton. Retrieved July 11, 2020 from: https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/2473.
Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly Use of Social Media and Altmetrics: A Review of the Literature. *Journal of the Association for Information Science and Technology*, 68 (9), pp. 2037–62.
van Leeuwen, T. N., Moed, H. F., 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. 335–46.
van Raan, A. F. (2005). Fatal Attraction: Conceptual and Methodological Problems in the Ranking of Universities by Bibliometric Methods, *Scientometrics*, 62 (1), pp. 133–43.
Vessuri, H., Guédon, J. C., & Cetto, A. M. (2014). Excellence or Quality? Impact of the Current Competition Regime on Science and Scientific Publishing in Latin America and Its Implications for Development. *Current Sociology*, 62 (5), pp. 647–65.
Visser, M., van Eck, N. J., & Waltman, L. (2020). Large-Scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *arXiv preprint*, 2005, 10732.
Waltman, L., & van Eck, N. J. (2016). The Need for Contextualized Scientometric Analysis: An Opinion Paper. In Ràfols, I., Molas-Gallart, J., Castro-Martinez, E., & Woolley, R. (eds). 21st Conference on Science and Technology Indicators, València, September 14–16, pp. 541–9. Retrieved August 17, 2020 from: http://ocs.editorial.upv.es/index.php/STI2016/STI2016/paper/viewFile/4543/2327.
Wang, Q., & Schneider, J. W. (2020). Consistency and the Validity of Interdisciplinary Measures. *Quantitative Science Studies*, 1 (1), pp. 239–63.
Weinberg, A. M. (1962). Criteria for Scientific Choice. *Minerva*, 1 (2), pp. 159–71.
Weingart, P. (2005). Impact of Bibliometrics upon the Science System: Inadvertent Consequences? *Scientometrics*, 62 (1), pp. 117–31.
Wilsdon, J., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., & Wouters, P. F. (2017). Next-Generation Metrics: Responsible Metrics and Evaluation for Open Science. Report of the European Commission Expert Group on Altmetrics. Retrieved June

22, 2020 from: https://op.europa.eu/en/publication-detail/-/publication/b858d952-0a19-11e7-8a35-01aa75ed71a1/language-en/format-PDF.

Wouters, P. (2014). The Citation: From Culture to Infrastructure. In Cronin, B., & Sugimoto, C. (eds). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press.

Wouters, P., Glänzel, W., Gläser, J., & Ràfols, I. (2013). The Dilemmas of Performance Indicators of Individual Researchers: An Urgent Debate in Bibliometrics. *ISSI Newsletter*, 9 (3), pp. 48–53.

Wouters, P., Ràfols, I., Oancea, A., Kamerlin, L., Holbrook, B., & Jacob, M. (2019a). *Indicator Frameworks for Fostering Open Knowledge Practices in Science and Scholarship* (Independent expert report for the European Commission). Retrieved July 11, 2020 from: https://op.europa.eu/en/publication-detail/-/publication/b69944d4-01f3-11ea-8c1f-01aa75ed71a1/language-en/format-PDF/source-108756824.

Wouters, P., Zahedi, Z., & Costas, R. (2019b). Social Media Metrics for New Research Evaluation. In Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (eds). *Springer Handbook of Science and Technology Indicators*. New York: Springer.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of Citation Performance and Excellence Measures: From Cross-Field to Cross-Scale Effects of Field-Normalisation. *Scientometrics*, 63 (2), pp. 373–401.