Defining and Measuring Research Impact in the Humanities, Social Sciences and Creative Arts in the Digital Age†

Tim Kenyon

Department of Philosophy, University of Waterloo, Waterloo, ON N2L 3G1, Canada, <tkenyon@uwaterloo.ca>

Tim Kenyon is Associate Dean of Arts (Research) and a professor in the Department of Philosophy at the University of Waterloo. His research focus includes issues in the philosophy of language, the epistemology of testimony, and critical thinking—including critical thinking about measures and rankings.

Tim Kenyon. **Defining and Measuring Research Impact in the Humanities, Social Sciences and Creative Arts in the Digital Age.** *Knowledge Organization. 41(3)*, 249-257. 16 references.

Abstract: There are powerful reasons for and against researchers taking the lead in formulating research impact measures for disciplines in the humanities, social sciences and creative arts (HSSCA). On balance, the reasons in favour are stronger, not least because such measures are otherwise apt to be formulated badly by those with little expertise. This invites us to inquire about the sorts of measures would best apply to HSSCA disciplines (among others), and whether some of the more popular impact measures, such as citation indices, really are reasonable indicators of impact or quality in these domains. It also raises questions about how burgeoning modes of research, knowledge mobilization, and impact tracking in the digital domain play into HSSCA research measures. On reflection, empirically adequate and arithmetically meaningful HSSCA impact measures will be pluralistic, non-reductive, and highly context-dependent; they are unlikely to lend themselves to the current pseudoscience of single-dimensional ordinal rankings between research institutions. Nevertheless they may support comparisons of interesting sorts, and enable assessments for accountability and planning purposes.

† A version of this paper was presented at the 2013 World Social Science Forum, whose anonymous referees provided useful feedback. For helpful input at various stages of this work, the author wishes to thank Jana Carson, Marc Couture, David DeVidi, Susie Gooch, Marie-Josée Legault, Gayle MacDonald, Bruce Muirhead, Tamer Oszu, Douglas Peers, and an anonymous reviewer from this journal. This work was supported by the Faculty of Arts, University of Waterloo.

Received 22 January 2014; Revised 26 February 2014; Accepted 26 February 2014

Keywords: research, impact, measures, citation, disciplines, researchers

1.0 Introduction

This paper will map out some conceptual issues bearing on the measurement of research impact in the humanities, social sciences, and fine and creative arts (HSSCA). I begin by canvassing some fairly general and discipline-independent reasons for and against the project of developing definitions and measures of research impact, before weighing these sets of reasons against each other. My conclusion, on balance, is that it is better to develop such measures, and that researchers in HSSCA fields themselves should take the lead in constructing and collectively endorsing the measures that will be used to characterize research impact. This raises the question of what such measures might look like, and why HSSCA disciplines are unlikely to be represented by some extant metrics that are increasingly employed to characterize research impact in the science, technology, engineering and mathematics (STEM) disciplines. It also implicates some respects in which digital scholarship and digital dissemination of scholarship offer both new opportunities and new challenges to the assessment of research impact.

These remarks focus on the most immediate forms of impact that research is commonly intended to achieve, and those forms deemed most directly to indicate academic strength, rather than considering impacts that might arise "downstream" of those immediate forms. For example, it is a virtual certainty that research conducted in HSSCA disciplines informs undergraduate teaching in those disciplines, and that the effects of that teaching are manifest in many significant economic, social, cultural, and political effects over the long term and at the population level. I will not consider impacts of these more distal sorts, which, as the London School of Economics Public Policy Group observes, might not yet be measurable to any useful degree of confidence or precision (LSE Public Policy Group 2011, 19)(although see Bornmann and Marx (2014) for a proposed scheme of societal research impact). Instead I will limit my remarks to issues arising with respect to the primary impacts of research. My aim is not to provide an exhaustive analysis of any of these issues, but rather to sketch their interrelations, and to motivate a way of thinking about research impact measures. To that end, I conclude by considering some plausible desiderata and constraints on HSSCA research metrics, given the issues described in these remarks.

2.0 Reasons to want explicit definitions and measures of research impact

Clear definitions and measures of research impact presumably share the virtues that accrue to the clarity and applicability of operational notions in general. The hope is that they will enable informative and longitudinally applicable self-monitoring by institutions, by academic units within institutions, and by disciplinary societies. Moreover, these tools would facilitate evidence-based strategic planning and resource allocation within and across institutions. And they would lend themselves to public accountability as well, an increasingly important request of public (or publicly funded) institutions of all sorts.

The allusion to explicitness is a critical element of the argument in favor of definitions and measures, since it highlights the respects in which inexplicit measurement is already practically ubiquitous in research culture. Discussions of research metrics sometimes do not attend to this fact, or fail to appreciate its implications. "Not everything that counts can be counted," begins a Stefan Collini essay critical of research bibliometry (2012, 120); but while Collini's objections to various particular research impact metrics are weighty, the aphorism is misguided with respect to counting. Used in this way, it suggests that there is some sort of fundamental category mistake involved in applying numbers to research and researchers. In practice, though, even the most mathophobic academic researchers already assess the relative quality and impact of other people's research all the time, if perhaps implicitly. These assessments are presupposed not just in obvious processes, like annual performance appraisals for research faculty, but in a wide range of other value-laden decisions. These include such choices as where researchers apply for jobs, where they try to place their students in graduate programs or junior faculty positions, and whom they invite to participate in conferences and colloquia. Even if one thought that each such evaluation was relativized to its specific purposes and contexts, still, within each context, researchers seem to have no problem doing things like advising a bright undergraduate student to apply to graduate program A as a number one choice, and graduate program B as a number two choice. In this light, the question isn't really whether to evaluate and rank research. It's whether to evaluate research via explicit methods.

Yet implicit evaluations and comparative judgments without explicit criteria are fertile ground for inconsistency, arbitrariness and bias across a wide class of domains (e.g., Greenwald and Kreiger 2006; Uhlmann and Cohen 2005). If research impact assessments are expressed only implicitly in academic behavioural choices of the kinds mentioned, there is no clear way of critically or constructively engaging the assessments—neither to endorse them meaningfully nor to correct them. By making research impact judgments explicit, and basing them on articulated principles, we enhance their clarity and fairness. In sum, while some of the reasons for wanting to develop research metrics and definitions implicate relatively recent pressures on researchers to explain and justify their work, in part the aim may simply be to do more rigorously and responsibly things that have long been done anyhow.

3.0 Reasons not to want definitions and measures of research impact

Analytic and conceptual tools of the sort contemplated here are plausibly on the horns of a dilemma. The measures will be misleading, pernicious, and inimical to accountability if they are badly formulated or excessively coarse-grained instruments. But they will be hard to interpret, hard to communicate to stakeholders, and hard to act upon if they are complex, nuanced and genuinely sensitive to the phenomena. Furthermore, even empirically sound measures may be misapplied, and put to uses that are harmful to the academy—or to public discourse about the academy.

For example, one way that such harms could arise is through crass comparisons of measures of research impact, even individually well-founded measures, across incommensurable categories. Different disciplines and different forms of impact might not bear out any very clear comparison of research quality or quantity. If one were confident that this incommensurability would not prevent unbridled attempts to make such comparisons anyhow, then one would justifiably be cautious about inviting the comparisons by generating the measures in the first place.

A related concern is that measures of research impact will contribute to the obsession with ordinal rankings of institutions among media, public officials, and even post-secondary education (PSE) administrative leaders around the world. As one such ranking exercise notes in its explanatory materials (van Vught and Ziegel 2011, 24):

> In university rankings … there is no scientific theory of what is 'the best university,' … no officially recognised bodies that are accepted as authorities that may define the rules of the game. There is no understanding, in other words, that e.g. the Shanghai ranking is simply a game that is as different from the Times Higher game as rugby is from football .... The issue with the some of the current university rankings is that they tend to be presented as if their collection of indicators did reflect the quality of the institution; they have the pretension, in that sense, of being guided by a (non-existent) theory of the quality of higher education.

It is not the purpose of this paper to debunk the innumerate and analytically vacuous university ranking schemes currently in vogue. It suffices to note that, to only slightly varying degrees, such rankings are arbitrary and obscurantist, and heavily laden with hidden unjustified value-judgments; they convey far more noise than signal. Moreover, virtually without exception they ignore or denigrate HSSCA disciplines in comparison to STEM disciplines. It is not an idle fear that HSSCA researchers would be fashioning their own noose by producing measures of research impact that would be abused in such rankings.

4.0 Weighing reasons

How, then, do these countervailing sets of reasons compare? In a sense, the reasons in favour of formulating research measures win out largely because the reasons against doing so have been overtaken by events. For research impact measures are coming, one way or another. Whether they are formulated, tested and implemented by HSSCA researchers or by, say, a private consultant working under contract to a regional education oversight body may well be the only practical question to settle. And if it is true that even well-designed measures can be used in misleading ways, or be hard to implement, nevertheless ill-designed measures and definitions are practically certain to be misleading and pernicious. Plausibly, ill-designed measures of research impact lend themselves far more to inappropriate comparisons, and other crass and silly uses. Definitions and measures of HSSCA research impact that are formulated from outside the community of HSSCA researchers are especially likely to be ill-designed.

These considerations lead me to conclude that generating measures and definitions of research impact is something that the community of scholars and researchers in HSSCA fields should take on themselves via grassroots initiatives. Cautious optimism about this prospect might be drawn from existing efforts, on both sides of the rather imprecise STEM-HSSCA divide, to advocate for appropriately nuanced and researcher-informed measures of research quality and impact. These efforts include recent explicit dialogue around humanities research measures in the Netherlands (Royal Netherlands Academy of Arts and Sciences 2011), and DORA, the Declaration on Research Assessment, spearheaded by the American Society for Cell Biology (2013).

5.0 Appropriate measures of impact: what's so special about HSSCA?

The need for breadth in the evaluation of research impact is widely discussed—Holbrook et al. (2013) sketch as many as 56 possible measures—but is far less honoured in practice. In many PSE institutions, especially where STEM disciplines are concerned, citation-count indices and other automated bibliometrics have increasingly been used as primary research quality and impact measures. The most coarse-grained version of such a metric is simply the total number of citations made to the work of a researcher, or to the work of all the researchers in an academic unit or institution. Of the somewhat more refined measures at the individual researcher level, one of the most heavily employed citation-count metrics is the Hirsch index, or *h-index*. For a given researcher, this index is the greatest number h such that at least h of the researcher's published articles have been cited at least h times in other published work. While it fails to convey information about the amount and impact of a researcher's scholarship having fewer than h citations, the h-index at least strikes some sort of balance between total number of citations and amount of published research at or above the h-number. Hence it expresses information that a total citation count would not.

What sort of information do citation counts and indices provide, in the broadest terms? They may be understood as proxies for research impact in the first instance, and via research impact, indirect proxies for research quality. Irrespective of discipline, their informativeness in these roles depends on empirical assumptions that can be quite fragile. Citation impact and research quality are of course very different things, the links between which can be obscure (Mryglod et al. 2013). But even the connection between impact and citation is profoundly complicated. Some of the factors discussed in the following section provide reason to doubt that real intellectual impact implies measurable literature citations, in many contexts, while other analyses suggest that the converse is equally dubious. That is, even extensive citation need not reliably signify a real intellectual impact, inasmuch as analyses indicate an alarming degree of irrelevant citation, and identical miscitations that propagate through some literatures; it is probable that citations are often being copied and pasted without the papers themselves being read (Todd and Ladle 2008; Simkin and Roychowdhury 2003). Jointly these reflections suggest that, even in the most favourable contexts of application, citation indices convey less information about impact than one might otherwise suppose.

To be sure, not all STEM disciplines in all parts of the globe have resigned themselves to a reliance on citation indices as the main proxy for research impact, still less research quality, while some social sciences in some places have embraced citation counts to a significant degree. So the STEM-HSSCA divide is substantially one of rhetorical convenience rather than of great precision. But with this caveat in place, the main point is fairly straightforward: even if we were to grant that measures like the h-index are somewhat useful proxies for research impact in some disciplines, chiefly on the STEM side of the academy, they are unlikely to capture impact for many HSSCA disciplines.

The most immediately obvious problem, but to my mind the least significant, is the difficulty of counting citations of published research in HSSCA disciplines. The databases and search engines currently used to index citations tend to under-count or entirely overlook the sorts of citations that are particularly characteristic of HSSCA research and publication culture: those appearing in books and in conference proceedings, for example. On one hand, this entirely disqualifies the use of citation indices as meaningful indicators of research impact for most HSSCA disciplines as things currently stand. On the other hand, it is the least interesting form of the problem, since it can and presumably will be resolved technologically, through more thorough data scanning, and better citation-crawling and counting software.

A more interesting and difficult problem, because it involves fundamental definitions, is the question of what counts as a citation for the purpose of calculating an h-number. For many HSSCA disciplines, collaborative input and influence on published research is not formalized through a relatively long list of co-authors, as in some STEM disciplines. Indications of such collaboration may be left implicit, or may be flagged in such forms as acknowledgements and thanks in the footnotes or prefaces of articles and books. This means that, in many HSSCA fields, citations to published work (or its analogues) are effectively citations to work influenced by researchers who are not named as authors on that work. So correlating citations with authorship is a far less reliable means of measuring scholarly influence and impact in many HSSCA disciplines than in many STEM disciplines. This leaves open the prospect of tracking citations to non-authorial traces of influence in published scholarship, such as acknowledgements in footnotes or prefaces; but no extant citation index system has made a move towards collecting this sort of information. Still other countable-but-uncounted phenomena, such as the use of articles or books as readings in research workshops or graduate seminar classes, might also be analogues of citation in other disciplines.

The question of what to count as a citation is of a piece with the still more acute question of what to count as research output. Books, conference presentations, policy-writing, legal opinions, creative performances and gallery shows may all be expressions of research output in HSSCA disciplines. Here it is especially clear that the problem isn't how to enumerate these things; it's what to count, and when to count it, and how much it counts for. These questions are *not* settled by software, no matter how greatly improved. They hinge on the intellectual values and academic practices specific to disciplines and sub-disciplines. Nor is the point exclusively one of HSSCA disciplines; it extends to STEM disciplines in which primary research outcomes may include such elements as patents, or changes to professional practices. There is no obvious reason to expect commensurability of comparison for these research outputs with the outputs and impacts characteristic of other disciplines.

A recent exchange between economist Richard Layard and philosopher Julian Baggini illustrates how hard it can be to bear this crucial pluralism point in mind. In an interview-style debate over the use of wellbeing indices in planning and resource allocation, Layard reacts to Baggini's suggestion that a broader notion of public agreement on multiple distinct priorities would be better than a common index-based approach. Layard (2012) objects: "But not everybody agrees that the same things are important .… Unless you have a single metric you cannot have a rational debate about priorities." In this quotation we see the problem in miniature: if we confront the genuine complexity of potentially incommensurable measures, then making judgments among different dimensions of comparison is hard. Whereas, if we stipulate a single metric, it's much more feasible to choose some outcomes as rational winners.

The problem, of course, and the reason why such blunt statements as Layard's are rare, is that this seeming rationality is spurious in cases where we know that the underlying reality is a mass of unsettled questions about values and best practices in measurement. Single-dimensional metrics generate fake clarity while obscuring actual complexity in such cases; the resulting judgments may permit fully ordered rankings that have the veneer of rigor, yet be little better than pseudo-scientific claptrap. The same worry arises in the case of research impact measures, once we take seriously that there are many kinds of impact and many proxies for the different kinds. The assumption that a common metric can be used to represent all these measures certainly holds out the promise of simplified reasoning and clear comparisons between fields, sub-fields, academic units, or entire institutions. The problem is that any such metric is likely to encode empirical and value-laden assumptions about how, and how much, to count each different component of the overall metric. This leaves us with our original problem of implicit, hidden biases; except now we've gone through the motions of using explicit criteria and are apt to be under an illusion of objectivity. The net progress over having no definitions or measures of impact may well be less than zero. In short, it is best to exercise great caution with the assumption that the plurality of impact types characteristic of HSSCA disciplines can be reduced to a single common metric.

A further issue distinguishing many HSSCA fields from many STEM fields is that of arithmetic comparability of citation counts and publication/output counts across disciplines. Both publication rates and citation practices vary considerably in the academy as matters of disciplinary and sub-disciplinary culture. The fields of medicine, most physical and biological sciences, and some social sciences have relatively high citation cultures; that is, disciplinary practices involve citing many other papers in one's own published research. Many humanities, fine and creative arts, and some mathematical disciplines have lower citation cultures (Expert Group on Assessment of University-Based Research 2010, 37):

> Publication and citation practices differ significantly from one discipline to another. In some fields, researchers may publish several research articles per year, while in other fields one monograph every 5 years may be appropriate. Citation frequencies also differ across disciplines. This has direct consequences for the journal impact factors published, for example, by Thomson Reuters in its Journal Citation Reports. In mathematics, a journal impact factor of 1.0 is high whereas in biochemistry journals with an impact factor of 1.0 is in the lower range. In the social sciences and humanities, journals tend to have impact factors below 1.0.

The variations in citation practices between fields (and between subfields) can in some cases be mitigated through statistical normalization. Roughly, this means scaling the numbers that characterize a discipline's citation culture so that the average number of citations (and, as far as possible, the distribution of citations relative to the average) is common across disciplines. In theory, this approach has the potential to permit meaningful comparisons across disciplines or sub-disciplines. Whether it works in practice depends on where and how one tries to apply it.

Normalizing for sub-disciplinary variations will work best for disciplines having a high citation culture in the core of the discipline (a high mean citation rate) and relatively small differences in citation practices associated with the sub-disciplines—either those falling substantially within the overall discipline, or those associated with interdisciplinary studies. However, disciplines with low citation cultures, and having high variability associated with sub-disciplinary and interdisciplinary work, will tend to make meaningful normalization difficult over the shorter term. The effects of the high-citation outliers will be disproportionately large, and the low mean citation rate, being bounded by zero, will generate a relatively narrow curve apart from the outliers. If the mean citation rate for journal articles in a discipline after (say) five years is 10, then a higher impact article might have 15 citations and a lower impact article might have 5. But if the 5-year mean citation rate in a field is between 1 and 2, there is effectively no way for a particular article, researcher, or academic unit to come in below that mean in a way that encodes interesting information about research impact, relative to the standards of the field. Much higher rates of citation for a particular researcher or article, by contrast, may simply indicate a readership somewhat outside the core demographic of the discipline.

This does not mean that the arithmetical process of normalization will somehow be impossible to execute in such cases. It is always possible to plug numbers into a formula and get numbers out. The concern is whether this will generate very meaningful results for at least many HSSCA fields, and a few STEM disciplines a well, such as pure mathematics (Bensman, Smolinski and Pudovkin 2010). A low core citation culture with high variability at the interdisciplinary margins is a relatively common feature among humanities fields in particular, suggesting that citation indices, however statistically polished, are unsuited to enable meaningful comparisons between these disciplines and others.

6.0 How does digital scholarship change the game?

The interactions of digital technology and culture with academic research add complexity to these issues in at least two key respects. In particular, digital considerations greatly complicate two basic questions we have already considered: What counts as an output? And what counts as an impact?

Owing to processes like faculty annual performance reviews, the demarcation between research production and research dissemination is already something of a vexed question within academic research institutions. Whether activities like giving public talks ought to count as research output (a quality-controlled placement of research results in an academically endorsed venue) or outreach (an informal discussion of research primarily among non-specialists), is a question asked and tentatively answered in different ways across the academy.

The differences in views on this question represent variations in institutional and disciplinary culture, but they represent also the extent to which the heterogeneity of the phenomena may be overlooked or oversimplified. This becomes particularly apparent when one considers digital domains. For example, simply to ask whether academic blogging counts as output or ("merely") as outreach (assuming that this general distinction really is well-defined) is to make some powerful assumptions about the unity of *blogging* as a category. In fact this category is so inclusive as to be of dubious value if employed without considerable qualifications. A blog can be a repository of researcher's thoughts of dubious relation to their expertise; it can be a channel for communicating independently published or validated scholarly results; it can be a group-moderated source of expert analysis that recapitulates in miniature the peer-review processes characteristic of the most traditionally prestigious research publication venues. In principle, a single blog could contain each of these elements over time, or in distinct forums under a single website name and URL.

It is increasingly recognized that digital forms of research output challenge and disrupt some of the chief means of traditionally recognizing higher-quality scholarship. Of course the meaningful aspect of peer review for research was never directly effected by having a major academic press or society produce a periodical in print, for which libraries paid subscription fees and individual researchers perhaps paid publication fees. But those features of the process were, and to some extent remain, hallmarks that are contingently associated with research quality control. They tend to indicate that some key conditions justifying the default trust of scholarship have been met. By contrast, digital venues for the presentation of research hold out the promise of open access to scholarly work, and of greater public discussion of research. But they also subvert the easy associative shortcuts that both researchers and the public have used to recognize peer-reviewed scholarship. This forces entire communities to make considered judgments about research provenance and research credibility, where mere feature-recognition used to suffice. These judgments are informed by fairly basic epistemic and value-theoretic commitments that can be hard to make explicit or to self-diagnose.

Similar digital complexities arise at the level of research impact, where they have occasioned discussion in part under label of "alt-metrics." It is unclear how to interpret, trust and weigh such potential impact measures as website hits, downloads, and searches. But if citation counts were already unhelpful as impact measures in the HSSCA fields owing in part to the wide range of impacts that research can have, this problem becomes far more complicated with the number and kinds of impact that digital outreach affords contemporary researchers and audiences. Digital dissemination may well fail to distinguish between the access that laypersons, policy-makers, influential public or business agents, or other academic researchers have made of research available online. Hence the chain of dissemination for research results, and the occasions of influence on opinions, actions and policies at all levels, have become increasingly difficult to detect and record, and harder still to quantify.

HSSCA research of many kinds will be particularly impacted by such considerations: creative work that is intended to be viewed or heard as performance, or academic research dealing with socially pressing or sensitive topics, will find a wide audience in the digital domain. Yet these issues of digital dissemination will arise for all academic research, STEM disciplines included. In this respect, the game-changing effects of digital technology and culture deserve special emphasis for HSSCA in part because some STEM disciplines have already nailed citation count indices to their masts as the chief proxies for research impact.

7.0 Desiderata on sound definitions and measures of HSSCA research impact

I will not close these remarks with a proposed definition of research impact; my contention is in part that no single such definition will be substantive while yet being broad enough to capture the range of discipline-specific forms of impact. Rather I will close by proposing some working principles for the construction of those definitions. These are largely intended to address or accommodate the factors considered in the foregoing remarks; whether they are the best ways of addressing those factors is not something I will argue here.

7.1 On a sound definition and measurement of research impact, good research is what researchers say good research is

The challenges canvassed in the previous sections combine to underscore the need for measures of research impact that are driven by the rich variegation of actual research, scholarly, and performative/creative practices in HSSCA disciplines. That is, discipline-specificity is a constraint on research impact measures for HSSCA fields. In practical terms, this means that discipline-based researchers are the right people to take the lead in formulating, testing, and revising discipline-based definitions and measures research impact.

This is not to say that the researchers in a field automatically have sole expertise on how to make explicit the justifiable core of their implicit reasoning and practices, when it comes to evaluating and comparing research impact. The guidance of such a reflective process may well be an independent expertise, brought into the process by facilitators from outside the discipline. But the knowledge itself, both in formulating impact measures initially and in seeking a reflective equilibrium over time in applying those measures, rests with experts in the discipline. The model of non-experts determining what experts should regard as research excellence and influence in their field is not viable. It is likely to misrepresent fields, and unlikely to secure buy-in from the core constituency of researchers.

What would such a grassroots approach look like in implementation? This is a question admitting of many answers, but one plausible suggestion for a starting point is to have a trusted neutral transdisciplinary academic body, national or international, facilitating and coordinating the efforts of various professional academic societies in HSSCA fields.

7.2 A sound definition and approach to measuring research impact enables comparisons

Another constraint concerns the prospects for non-trivial aggregation of measurement results within at least some cohort for comparative purposes. The research impact measures that a discipline settles upon should not be formulated in a way entailing that every researcher, every department, or every program is a singleton set. It will also be important to permit meaningful comparisons within aggregates: in short, everyone can't be tied for best along every dimension of comparison.

The idea here is just that academics within a single discipline do in fact make research-based comparisons: researcher to researcher, department to department, subfield to sub-field. If research impact is defined in a way that makes every such comparison fallacious, then we will have refined a notion of impact that fails to make contact with the actual uses for which researchers used the pretheoretic notion in the first place. Maybe all such uses are unjustifiable; we can't rule that out as something to be discovered. But as a working principle we should assume that there is a recoverable core of existing practice that a good definition can capture.

7.3 On a sound definition and measurement of research impact, comparisons are context-sensitive

From sports league tables to Top Ten lists, the very idea of comparison popularly carries with it the assumption of straightforward ranking from best to worst. For complex multidimensional phenomena, however, comparability does not carry this implication. Such a ranking is perhaps the least revealing and the most misleading form of comparison possible, for reasons discussed earlier.

Yet the fact that there is no single overall answer to a question like, "What is the best car?" doesn't mean that it is irrational to comparison-shop when buying a car. It just means that the question becomes a meaningful one in the context of assumptions about the kind of car in question, and what one wants from a car. Similarly, the comparisons afforded by a good definition of research impact in HSSCA will not purport to settle a general question like, "What paper, researcher or department has the biggest research impact?" It will, however, facilitate comparisons when appropriate contextual information and norms are factored in.

In principle this could be said even of simplistic citation counting; it's just that a unitary metric like that has few degrees of freedom to be influenced by context. But in the HSSCA case, the plurality of potential research impact types means that the explanatory needs or interests of the situation can have a powerful effect on which impacts count as the most important in that context. Returning to the analogy: you might rent a small car having the best fuel economy one weekend, and rent the largest van on the lot to move furniture the following weekend. In each case the context settles a genuine question of what counts as the best vehicle at that time, without carrying any suggestion that there must be an answer to the question of which vehicle is the best, period. Similarly, impact comparisons are inevitably informed by context-specific valuations of impact-kinds. Just as in the vehicle case, there is no reason to expect that such comparisons make sense beyond the quite localized contexts that informed them.

7.4 For the sound definition and measurement of research impact, there is an explicit temporal variable linked to the characteristics of the research culture in question

A child of my acquaintance, having been asked to boil some water, was observed to fill a pot with cold water, place it on the stove, and immediately announce, "It's not working!" Like water, research too boils on a schedule that does not necessarily accommodate itself to the amount of time a particular analyst wishes to wait for it.

Politicians are typically elected for terms of three to six years, most of them being four or five years long, while senior university administrators are generally appointed for terms of a similar length. It is not, I submit, a coincidence that researchers both singly and in aggregate find themselves asked to provide evidence of research impact carved up into temporal swathes that generate results within electoral and contractual employment cycles; people want to know what has been accomplished on their watch. But should we expect the time frames appropriate to the analysis of research impact to match up with time frames of administrative convenience? We do not expect to see redwoods grow to maturity in five years; we would not use a two-year window to measure the success of a sea-turtle breeding program. That might be inconvenient for people wishing to study those phenomena, but there it is.

So the question is: what is the right length of time over which to measure research impact in a particular HSSCA field, if we expect a meaningful answer? In part this is an empirical question about how long it can take for research to have significant influence in a particular discipline. But in part it is a question of values; it forces us also to ask how long we are prepared to wait in order to distinguish more impactful research from less. Neither the empirical nor the normative questions can be engaged unless we make explicit that our evaluative time frame is a choice. If that choice is made arbitrarily—if we simply stipulate a two-year, ten-year, or rolling seven-year window of impact analysis, we once again indulge in a false clarity. A good definition of research impact will make it hard to do this, by making the temporal measurement aspect an explicit choice to be justified.

7.5 The sound definition and measurement of research impact lends itself to uptake and ongoing facilitation

Perhaps the most compelling objection canvassed earlier to the project at hand was this: No matter how carefully qualified and appropriately nuanced a definition or set of metrics may be, somewhere there is a decision-maker waiting to use it as a hammer. This is not a prospect that has to be regarded passively, however. A good approach to research impact analysis will craft the definitions and methods in a way that explicitly invites or requires expertly guided facilitation. The guided expert interpretation of measures of research impact should be part of the model, as others have noted as well (van Leeuwen 2007, 105):

> An important issue related to the discussed bibliometric research performance assessment model, its understanding, and the interpretation of bibliometric analyses in general, is the role of the person, the bibliometrician. A bibliometric researcher can and should inform the users of bibliometric data on the advantages and disadvantages of this type of study, through dialog. By answering questions regarding the function and goal of a proposed bibliometric study, the bibliometrician can guide the initiator of any bibliometric study in the direction that leads to the application of the most appropriate approach, and the related techniques.

Of course this is no guarantee against misuse, either willfully or through ignorance. But the more explicit the cautionary notes requiring expert HSSCA facilitation, the fewer misuses we may expect—and the greater recourse researchers will have to remedy those misuses, since they will be such clear violations of the definition and methodology.

Perhaps this is the right note on which to close these remarks: not only should HSSCA scholars take the lead in formulating definitions and measures of research impact, but they should formulate them with the explicit aim of remaining involved in the implementation of those measures over the long term. Writing a handbook of research impact assessment with no plan beyond placing it in the hands of policy-makers and resource-allocators is surely both a waste of time and an abrogation of responsibility. Ultimately, the entire proposal can be understood as an expression of commitment to the very system that makes for truly innovative and progressive research in the first place: collegial governance among researchers. When it comes to research impact assessment, collegial governance entails following through with guidance and advocacy on the use of assessment methods.

References

American Society for Cell Biology. 2013. *San Francisco declaration on research assessment.* Available http://am.ascb.org/dora/.

Expert Group on Assessment of University-Based Research. 2010. *Assessing Europe's university-based research.* Available http://ec.europa.eu/research/science-society/document_library/pdf_06/assessing-europe-university-based-research_en.pdf.

Bensman, Stephen J., Smolinski, Lawrence J. and Pudovkin, Alexander I. 2010. Mean citation rate per article in mathematics journals: differences from the scientific model. *Journal of the American Society for Information Science and Technology* 61: 1440-63.

Bornmann, Lutz and Marx, Werner. 2014. How should the societal impact of research be generated and measured? a proposal for a simple and practicable approach to allow interdisciplinary comparisons. *Scientometrics* 98: 211-9.

Collini, Stefan. 2012. *What are universities for?* London: Penguin.

Greenwald, Anthony G. and Kreiger, Linda Hamilton. 2006. Implicit bias: scientific foundations. *California law review* 94: 945-967.

Holbrook, J. Britt, Barr, Kelli R. and Brown, Keith Wayne. 2013. Research impact: we need negative metrics too. *Nature* 497: 439.

Layard, R. 2012. Interview by Susanna Rustin. Can happiness be measured? *The guardian* July 20 2012. <http://www.theguardian.com/commentisfree/2012/jul/20/wellbeing-index-happiness-julian-baggini> Accessed May 20 2013.LSE Public Policy Group. 2011.

Maximizing the impacts of your research: a handbook for social scientists: consultation draft 3. 2011. Available http://www.lse.ac.uk/government/research/resgroups/LSEPublicPolicy/Docs/LSE_Impact_Handbook_April_2011.pdf.

Mryglod, Olesya, Kenna, Ralph, Holovatch, Yu and Berche, Bertrand. 2013. Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics* 97: 767-77.

Royal Netherlands Academy of Arts and Sciences. 2011. *Quality indicators for research in the humanities*. Available https://www.knaw.nl/en/news/publications/quality-indicators-for-research-in-the-humanities.

Simkin, M.V. and Roychowdhury V.P. 2003. Read before you cite! *Complex systems* 14: 269-74.

Todd. Peter A. and Ladle, Richard J. 2008. Hidden dangers of a "citation culture." *Ethics in science and environmental politics* 8: 13-6.

Uhlmann, Eric Luis and Cohen, Geoffrey L. 2005. Constructed criteria: redefining merit to justify discrimination. *Psychological science* 16: 474-80.

van Leeuwen, Thed N. 2007. Modelling of bibliometric approaches and importance of output verification in research performance assessment. *Research evaluation* 16: 93–105.

van Vught, Frans and Ziegel, Frank. 2011. *Design and testing the feasibility of a multidimensional global university ranking: final report*. Consortium for Higher Education and Research Performance Assessment. Available http://www.ireg-observatory.org/pdf/u_multirank_final_report.pdf.