British Philosophical Association

Metrics-Based Research Assessment

This document is the BPA's response to HEFCE's recent call for evidence on the use of metrics in research assessment. We think that metrics are not a good or reliable measure of research quality in philosophy, at any level (the quality of individual researchers or pieces of research, of the research carried by a particular journal, the research carried out by a research group, institution, or nation). They are in particular no substitute for peer review. Research quality is distinct from research impact, and the BPA's view is that metrics should never be taken as a measure of impact either.

1. Context

On 3 April 2014, the Government asked HEFCE to undertake a review of the role of metrics in research assessment and management. This review will build on a previous pilot exercise in 2008/9 which concluded that citation information was not sufficiently robust to be used formulaically or as a primary indicator of quality but that there might be scope for it to inform and enhance processes of expert review. The new review will explore the current use of a range of metrics for research assessment, consider the robustness of metrics across different disciplines, and assess their potential contribution to the development of research excellence and impact.

The term 'metrics' is used to refer to range of different ways of coming up with and analysing quantitative information. Bibliometrics focuses on the quantitative analysis of information about publications (e.g. citation data), webometrics or cybermetrics studies the many measurable aspects of online items, and altmetrics looks at measurable aspects of social media activity. The HEFCE's call for evidence asks for contributions explaining which metrics (if any) are already used in assessing (formally or informally) research excellence and impact in different academic disciplines, and to identify which metrics should be used for making evaluations of research quality, research impact or any other aspects of assessment, such as research environment.

2. Metrics in use

In philosophy, research is assessed for a variety of reasons, including:

- it is the function of the REF (and previously the RAE) that research be evaluated
- for making judgements about publishing written work
- for making judgements about hiring, retention and promotion of staff
- for making decisions about awarding grants

Philosophical research quality is primarily assessed by 'peer review', which is to say that some philosophers learn about another philosopher's research, analyse it in the context of existing research, and make judgements about the quality of those ideas or the activities and practices that were involved in it. Philosophical research is an activity, one outcome of which is 'output' in the form of a publication. Other outcomes include verbal presentations, reports aimed at policymakers, better and more up-to-date teaching, and, importantly, a high level of expertise both in a particular area of philosophy and in the intellectual virtues of clarity, robust argument, critical analysis, and so

on. **Since philosophical research outcomes are not limited to published outputs, peer review is not limited to assessing publications**. Currently in peer review at all levels (REF, publication, hiring etc.), metrics play no part in assessing the quality of the research; judgements about quality are made on the basis of the content of the research, and not on the basis of any of its quantifiable features. Use of bibliometrics, such as citation-counts, H-indices and journal rankings are not employed to determine research quality, although they are widely but critically discussed.

Judgements about research quality are made and disseminated in many ways – they are not limited to formal discussion (and citation) of written work in a publication. Less formally, discussion of another philosopher's work in a public forum such as a conference or workshop, in academic publications, or in classrooms can help to shape, fix and share these judgements. Some processes are much more formal, such as the REF, or triple-blind peer review procedures employed at increasingly many academic journals. We see room for further improvement in each of these processes to remove the effects of implicit and explicit biases. But it remains the case that very few of the judgements of research quality which result are recorded in any form that can be adequately measured by metrics. So it would be a mistake to use metrics as a proxy for judgements of research quality. **The analysis of publication metadata (citation counts, journal ranking, journal impact factors), or of grant income, or webpage ranking and keyword searches, or frequency of invitations to speak can only ever represent a misleadingly narrow part of the activity of evaluating other philosophers' work**. The BPA have previously expressed strong reservations about the narrowness of HEFCE's understanding of what kinds of things constitute philosophical research, research 'outputs', and what 'impact' is supposed to involve.[1] These same kinds of mistakes are in danger of being repeated by the current review of metrics. On this point in particular, it is unclear what HEFCE's review of the role of metrics is attempting to do; is the review:

a) looking to determine whether using quantitative metrics can speedily and reliably replicate or otherwise capture the conclusions of longer, slower qualitative methods of judging research excellence? or

b) looking to determine whether metrics can be revisionary with regard to existing ideas of research excellence, by revealing peer review to be inaccurate when it disagrees with quantitative data?

The BPA think that there is no quantitative substitute for academic judgement of research quality, and that the important concept of peer-evaluated research excellence should not be replaced or supplanted by a distinct, quantitative measure.

3. Metrics and judgements about research quality

Various metrics may *appear* to be indicators of scholarly excellence, or may *seem* to capture (in some sense) qualitative judgements of scholarly excellence:

- research grants awarded
- numbers of research students and their completion rates
- publishing frequency
- rankings of the journals in which those publications occur
- citation rates / H-indices
- page ranking / webometrics
- social media activity / altmetrics

---

[1] BPA position paper on the assessment of impact in the REF (November 2009). Available online here: http://www.bpa.ac.uk/uploads/2009/11/bpa-impact-position-paper-301109.pdf

Not all of these metrics equally reflect prior judgements of quality by suitably qualified people; altmetrics and webometrics are particularly problematic in this regard. In contrast, bibliometric measures involving journal-ranking may seem to capture information concerning the outcome of a rigorous peer-review process used in publication. But in what follows we distinguish between two types of claim: that a high score on one of these metrics is sufficient for indicating quality research, and that a high score on one of these metrics is necessary for thinking that a piece of research is excellent.

Each of these metrics is problematic in at least three distinct ways.

1. They are unreliable as measures of research quality.
2. Reliance on these metrics in making judgements about research quality can be problematic for structural and political reasons.
3. Using these metrics to assess research quality can incentivise research activities which do not produce high quality work (see §5 below)

On the first point, we think that while bibliometrics, cybermetrics and altmetrics can successfully reveal patterns of certain types of activity, they cannot reliably reveal judgements of research quality. There may be many other reasons, besides research quality, why some published outputs, individual scholars, research groups or university departments attract more attention (citations, page counts, activity on social media) than others. **In general, because cybermetrics and altmetrics do not necessarily capture prior judgements by suitably qualified people, they should be considered highly unreliable**.

The HEFCE's Research Excellence Framework requires its panels of peer reviewers to use criteria of 'originality, significance and rigour' when evaluating the research quality of published outputs.[2] **None of these criteria are adequately captured by bibliometrics** (the analysis of quantitative information about publications); in what follows we focus on citation counts, journal ranking, and the award of research grants.

> 'Originality' and citation counts. Highly original work may not receive a sufficiently significant number of citations for a host of reasons, including: it may be published in an unusual or relatively inaccessible venue; it may be relevant to only a very small community of researchers working on a given area; it may be sufficiently radical to be perceived as marginal. A high citation-count is never necessary for a piece of work to be considered original. In addition, highly unoriginal work can receive a significant number of citations for similar reasons: it may simply be published in easily accessible places, or with broad appeal across many disciplines. If follows that a high citation-count is no guarantee whatsoever that a piece of work is original.
>
> 'Originality' and journal ranking. Highly original work may not be published in highly-ranked journals for a number of reasons: it may have a high level of technical rigour (formal notation and so on), it may be highly original in its level of generality or its narrowness of focus, or it may be sufficiently radical to be perceived as marginal, any of which might mean that the best place for its publication may not be among the highest-ranked journals. Being published in a highly-ranked journal is never necessary for a piece of work to be considered original.
>
> 'Originality' and the award of research grants. Unlike STEM subjects, and in common with most work in the Humanities, highly original research can be carried out without needing to

[2] Further details about assessment procedures can be found here: http://www.ref.ac.uk/panels/assessmentcriteriaandleveldefinitions/

apply for many research grants. Success at applying for funding is not necessary for a scholar (or their project) to be considered highly original.

'Rigorousness' and citation counts. High levels of technical rigour in philosophical publications can be a direct reason for its having fewer citations, since there are fewer journals (with smaller readerships) in which to publish such work. A high citation-count is never necessary for a piece of work to be considered rigorous. In contrast, careless or negligent work, once published, can be highly cited, simply because many scholars wish to point out its problems. A high citation-count is no guarantee whatsoever that a piece of work is rigorous.

'Rigorousness' and journal ranking. Highly rigorous work may not be published in highly-ranked journals precisely because it has a high level of technical rigour, and so is better suited to a journal with a narrower remit and (accordingly) a less prestigious ranking. It follows that being published in a highly-ranked journal is never necessary in order for a piece of work to be considered rigorous.

'Rigorousness' and the award of research grants. As above: highly rigorous work in philosophy can be accomplished without having to secure research grants. Success at applying for funding is not necessary for a scholar (or their project) to be considered highly rigorous. In addition, relatively less-rigorous research groups can afford to dedicate a lot more time (or research members) to applying for lots of awards, such that their rate of success can seem relatively high. It follows that success at applying for funding is not a good guarantee that a research group is rigorous, nor that the money which they have secured has been well spent.

'Significance' and citation counts. For many of the same reasons that are given above, the 'significance' of a piece of published work, scholar, research group or department is not measured by citation-counts: work may be highly cited because it is significant, but it doesn't follow that highly cited work is significant, or that less cited work is less significant.

'Significance' and journal ranking. Highly significant pieces of research can be published outside the top-ranking journals for a variety of reasons, including those already covered. In general, being published in a highly-ranked journal is never necessary in order for a piece of work to be considered significant.

'Significance' and the award of research grants. As above: highly significant philosophical research can be carried out without research grants, while at the same time its possible for research groups with a less significant agenda to apply for a wider variety of grants more frequently, such that their rate of success can appear relatively high. Work may be funded because it is significant, but it doesn't follow that highly-funded work is significant. Grant awards are neither necessary nor sufficient for thinking that philosophical work is significant.

There are other issues with the use of each of these metrics in philosophy. A problem with using journal ranking in philosophy in general is that **no single ranking of all philosophical journals will ever capture the relative importance of each journal to each philosophical sub-discipline**; as all appeals to a journal-ranking-score as a sufficient indicator of relative quality are only as reliable as the method of journal-ranking, this is a serious issue. Similarly, awards of research grants are not a sufficient indicator of quality since **not all grant-making trusts make awards on the basis of the judgements of suitably qualified people**, or on the basis of research quality at all. Philosophy as a discipline is one of the smallest in the Humanities, so **specific sub-disciplines can be very small indeed, nationally and internationally, and thus score very low on metrics such as citation-counting**. Furthermore, as already suggested,

metrics based on journal rankings and citations assume an excessively narrow conception of what kinds of things constitute philosophical research.

We think that attempting to measuring 'significance' or 'importance' quantitatively is independently problematic: patterns of activity around some scholars may be marking out academics who are one or more of the following: already established in academic circles; publicly visible and audible (for a wide variety of reasons, both good and bad); powerful and thought to be worth flattering; popular and thought to be worth knowing. This brings us to point 2 above.

We are concerned that many of these metrics may measure something like academic fame or influence. This is problematic because philosophy is dominated by white men in way that many other disciplines – particularly in the arts and humanities – are not. There are very few black and ethnic minority philosophers in the UK, although comparisons with the proportion of black and ethnic minority scholars in other disciplines have not yet been carried out. The Society for Women in Philosophy (SWIP) and the BPA have issued a joint report on the problematic underrepresentation of women in philosophy.[3] Our concern is that philosophers may be unfairly downgrading the work of, or wrongly ignoring or mistakenly failing to encounter, women philosophers or black or ethnic minority philosophers at all career stages. SWIP and the BPA are taking action to improve peer-review processes, and the procedures surround the hiring, retention and promotion of staff, precisely by drawing attention to this issue. But **any tendency to cite established authors, whether intentionally or a result of implicit biases, will unfairly discriminate against women philosophers or black or ethnic minority philosophers in metrics**. The same problems affect the other metrics: metrics which elide the differences between academic fame or influence and academic excellence will perpetuate the same structural injustices. Relatedly, in our view the use of metrics to measure research impact or excellence will discriminate against excellent philosophers working on less established areas of philosophy, or in particularly small sub-disciplines, since it can be harder (in various ways) for such scholars to achieve recognition or exposure for their work.

In summary, **we think that there is no good quantitative method for capturing the kinds of verdicts which result from peer review**. Of those under discussion, journal-rankings may seem to most-closely record some of the judgements that professional philosophers make about each others' work in peer-review. Our considered opinion is that there is no one journal-ranking which is able to do this while adequately representing the sub-disciplinary specialisms and the plurality of approaches which characterise academic philosophy. Appeal to journal-rankings places undue emphasis on an excessively narrow conception of what kinds of things constitute (high quality) philosophical research.

4. Metrics and judgements about research impact

The BPA think that the HEFCE guidelines on evaluating the 'impact' of philosophical research are insufficient. We have previously expressed many of these concerns in our position paper on the decision to include an evaluation of 'impact' in the REF.[4] This statement builds upon that previous statement, and updates our position:

- We are concerned by the prospect that metrics may be used to measure research impact.

---

[3] SWIP/BPA Report: Women in Philosophy in the UK (September 2011). Available online here: http://bpa.ac.uk/uploads/2011/02/BPA_Report_Women_In_Philosophy.pdf

[4] BPA position paper on the assessment of impact in the REF (November 2009). Available online here: http://www.bpa.ac.uk/uploads/2009/11/bpa-impact-position-paper-301109.pdf

- **Philosophical research is an activity which can make a difference both to society at large, and to the academic communities in which it flourishes. However, this difference is often long-term, unpredictable, and hard to quantify.** The history of the discipline shows many examples of works now universally recognised to be some of the most important philosophy ever written which 'fell dead-born from the press'.

- **There is a wide variety of ways through which philosophers' research can have a measurable impact,** including: verbal presentations; reports aimed at policymakers; better and more up-to-date teaching; making radio programmes and podcasts; writing or commenting on newspaper / magazine / blog articles; participating in or supporting informal discussion groups or activist networks; collaborations in both inter-disciplinary and non-academic forms, such as discussions with policymakers, scientists or school teachers; achieving a high level of expertise both in a particular area of philosophy and in the intellectual virtues of clarity, robust argument, critical analysis, and so on. **Few of these outcomes can be assessed using metrics, since their importance is not measured quantitively.**

- We think that there are ways that some research impact can be assessed (e.g. expert peer review), but we think that metrics are not a useful gauge of research impact, for parallel reasons to those given in §3:

    1. They are unreliable as measures of research impact: metadata reveal patterns of activity that are no substitute for qualitative judgements.

    2. Reliance on these metrics in making judgements about research impact – where those metrics are insensitive to the effects of academic fame or influence – will discriminate against women philosophers and black and ethnic minority philosophers, since they have a lower profile across the profession in the UK, so this can be problematic for structural and political reasons.

    3. Using these metrics to assess research quality can incentivise research activities which do not produce high quality work (see next section)

5. 'Gaming' and strategic use of metrics

There is little evidence of which we are aware that the behaviour of researchers, research managers or publishers has been affected by strategic use of metrics in philosophy, primarily because metrics have so far had no role to play in the assessment of research quality or impact in philosophy. However, on the basis of richly suggestive evidence arising from engagement with the REF, we believe that the introduction of metrics is likely to encourage strategic approaches to research which do not necessarily result in high quality work; **energy may be diverted away from the pursuit of excellent philosophical work (as currently understood) in order to attempt to produce outputs that score more highly on particular quantitative scales**.

For example, academics may be encouraged, tempted (whether consciously or not), or put under pressure to write something provocative and attention-seeking rather than something thoughtful, nuanced and well-supported. Just as making citations the principal measure of research quality effectively incentivises people to write such pieces, so the concern about the strategic use of metrics generalises to other metrics, for other forms of assessment. Making grant awards the principle measure of a research groups' excellence incentivises such groups to spend their awards paying for researchers to apply for more awards, and making web traffic a measure of a philosophy department's research impact incentivises those departments to spend more time generating web traffic than doing excellent philosophical work.