Better, fairer, more meaningful research evaluation – in seven hashtags

*Considering the future of research assessment, **Elizabeth Gadd** outlines how she believes research evaluation could be made better, fairer, and more meaningful. The resulting seven guiding principles, neatly framed as hashtags, range from understanding our responsibilities to researchers as people, through to ensuring our evaluations are a more formative process, offering valuable, constructive feedback.*

Imperial College recently held an event called "Mapping the Future of Research Assessment". It was a chance for Imperial College staff to consider how becoming a DORA (Declaration on Research Assessment) signatory would affect their approach to research evaluation and I was kindly invited to speak more broadly about the concept of responsible metrics. In putting together my talk, I found myself trying to articulate what guided me when I sought to do metrics responsibly, as well as how I'd like to see the world of research evaluation improve. What resulted was seven high-level concepts – principles, if you like – that, I believe, would make research evaluation better, fairer, and more meaningful. And as I spend way too much time on Twitter it seemed only natural to frame them in hashtags. A recording of the event is available and what follows is a transcript (almost) of that.

1. #PeopleArePrecious

I think any responsible research evaluation journey has to start by remembering who we are responsible *to*. As a human race, we are utterly bonkers if we do not prioritise creating the most fertile environment for the best, brightest, and most creative minds of our generation to do their best work; to save and enhance lives; to create a better future. Progress utterly depends on it. But I fear the opposite is happening. People are leaving higher education, just as they left primary and secondary education, due to crazy workload pressures, unrealistic expectations, poor pay, and weaponised metrics. Metrics of Mass Destruction. Unrealistic targets, unfair comparisons, blunt-instrument rankings, and narrowly focused indicators that don't serve everyone equally.

Whilst blood, sweat, and tears are finally being expended trying to create a more equal and diverse academic workplace, are we at the same time homogenising the workplace through metrics that don't give everyone the same chance to succeed or reward the right things? At the same time as paying lip service to openness, are we also closing down access to half the globe – who once couldn't afford to buy the "best" journals and now, in a cruel twist of fate, can't afford to publish in them either, and thus will never get the recognition of narrow-minded evaluation schemes reliant on journal-based metrics?

All people are precious. OK, I'm not saying they aren't also selfish, cruel, and bloody annoying at times. But as physicist Carl Sagan said: "Every one of us is, in the cosmic perspective, precious. If a human disagrees with you, let him live. In a hundred billion galaxies, you will not find another". Our best minds should be nurtured and supported, not unfairly measured and ranked.

2. #GetGranularityRight

In an effort to protect our precious people, I think we need to make sure that when we're doing research evaluation, we get the level of granularity right. Indeed, when research evaluation goes wrong, it's often because we measure at the wrong level of aggregation. The classic case is the use of Journal Impact Factors to measure individual authors or individual articles – something they weren't designed to do.

But it's not just Journal Impact Factors. We are running more and more research evaluation at the level of the individual researcher and setting targets at the level of the individual researcher, forgetting that in all STEM disciplines, and increasingly in AHSS disciplines, research is done in *teams*. It takes a team to win a research grant, do the research, write papers, and generate impact. So, I'd ask, where is the logic in extrapolating outcomes at the level of the individual? And where is the benefit?

Date originally posted: 2018-09-27

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/09/27/better-research-evaluation-in-seven-hashtags/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

In terms of grant income this can lead to individualism and competition – behaviours which are actually counter-productive when it comes to winning grants. In terms of papers and citations, we have to think about how fair and meaningful this is – especially in a world of increasingly hyper-authored papers. If I'm one of 1,500 authors on a paper, should I get the same credit for that paper as a single-author? Because in all current citation counting tools, I do. If we measured success at the level of the team – or the paper – this problem would go away.

At the other end of the spectrum of course, with national research assessment exercises like the REF, we are seeing evaluation at the level of whole departments and schools, again ignoring this important point that research is done by smaller units – research teams. Is it fair for a high-performing research group in a less-well performing school to fail to be rewarded for its efforts as a result of bundling?

It seems to me that getting your granularity right is critical to making research assessment fair and meaningful.

3. #InvestInPotential

I think another key way we can protect our people is to promote on potential. All research assessments (whether using metrics or peer review) are essentially backward-looking and based on past performance. We assume that because someone or something has performed in the past, it will do so in the future. But that is an assumption. And it is an assumption that serves the long-standing academic far better than the early-career researcher. If we think about the world of undergraduate recruitment, such assumptions serve the wealthy student who has had a better past than students from poorer backgrounds who may have had a rocky start. However, recent research at the University of York showed that students from disadvantaged backgrounds with lower grades (poor past performance) did just as well at medical school as advantaged students with higher grades.

Which leads us to ask – what if we all appointed based on potential? In one of the great recent DORA interviews, Sandra Schmid described the novel recruitment approach at UT Southwestern, in which they judge candidates not on previous publication performance but on future research plans. They're convinced this leads to more productive, long-term partnerships with their new recruits.

This applies at other levels of aggregation too. The Dutch Standard Evaluation Protocol (their national research evaluation exercise) has as one of its three main pillars the concept of "viability". Research groups are assessed on research quality, societal relevance, and their *future research plans*. How great is that? Imagine how the playing field would be levelled if our academic investments were based on evidence of potential, rather than solely on evidence of a past.

4. #MeasureWhatMatters

I'm not naïve enough to believe we can do away with all forms of measurement and nor should we. Of course we need to be accountable for the research pounds we are spending. We also need to ensure our decision-making processes are transparent and fair and this usually demands some objective measures to counter unconscious bias. But all too often we value what we can easily measure, rather than measuring what we actually value.

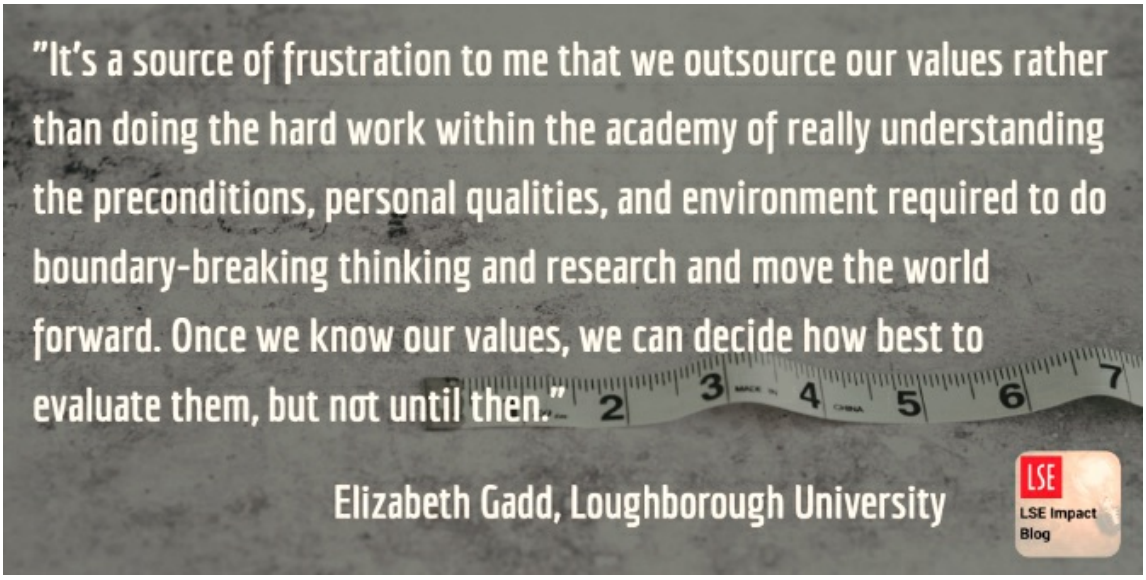
This is called the streetlight effect. We look for answers amongst bibliometric data because this is often the only "streetlight" we have, rather than articulating what we value and e-valu-ating accordingly. Because I'm not sure we always really know what we value. Indeed, too often we outsource our values to funders and rankers. How many institutions have KPIs around getting into the top 100 of Ranking X without any real understanding of what Ranking X actually measures? How many institutions have an open access policy which is an exact replica of the REF Open Access Policy without giving any thought to the standards of openness they might want to set for themselves?

It's a source of frustration to me that we outsource our values rather than doing the hard work within the academy of really understanding the preconditions, personal qualities, and environment required to do boundary-breaking thinking and research and move the world forward. Once we know our values, we can decide how best to evaluate them, but not until then.

Date originally posted: 2018-09-27

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/09/27/better-research-evaluation-in-seven-hashtags/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/



5. #RecognitionNotEvaluation

Having said we need to measure what matters and at the appropriate level of granularity, I guess I'd really like to see those measurements (and I include all forms of evidence here, including references, prizes, peer review, and so on) contributing to a *recognition* culture rather than an evaluation culture. Ask a room full of academics who'd like to be evaluated, and you won't see many volunteers. However, ask who'd like to be recognised and I'm pretty sure all hands would go up. Because evaluation is inherently critical and comparative, while recognition is inherently positive and celebratory.

I'm convinced that scholarship (and therefore society) would be best served by moving away from a hyper-competitive environment, fuelled by comparative evaluation, which is the birthplace of many mental health problems. Instead we should move towards an environment where, yes, all players have to meet a certain quality threshold (educational level; ability to teach, research, write), but beyond that, which also recognises that we all have different strengths that should be celebrated. It may be inspiring undergraduates, doing public engagement, entrepreneurship; I could go on. And of course, when we have a wider range of things to recognise, we need to recognise them all equally, and not pay lip service to holistic evaluation, and secretly only value pounds and publications.

6. #ProfilesNotRankings

One of the best ways we can move towards recognition and away from evaluation is to abolish rankings and instead create profiles. Trying to rank individual academics, research groups, or institutions is like trying to rank the fruit in your fruit bowl or your own children. You cannot rank multi-dimensional entities on a single scale labelled "best" or "top". And yet this is what we do, particularly with world university rankings. I've said before that any research proposal that sought to identify the "top" universities would be rejected for a badly formed research question. Top at what? And of course, not many of our rankings contain error bars which would expose how little confidence we can have that the university ranked #1 is significantly different to the one ranked #50. Indeed, if we did apply confidence intervals or other statistical techniques, we'd probably observe clusters of universities based on age, mission, and subject mix, and conclude that within those clusters they are pretty much of a muchness.

Once we put numbers on people, groups, or universities, we can't help ourselves but to line them all up in size order. Profiles would give us a much richer picture of strengths and weaknesses (and we all have both) and help us to make better-informed decisions based on the issue at hand.

7. #FormativeNotSummative

Date originally posted: 2018-09-27

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/09/27/better-research-evaluation-in-seven-hashtags/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

My final hashtag tries to get to the heart of what research evaluation is actually for. What is the point of research evaluation if it doesn't actually leave us with a better research system than the one we started with? And it's interesting to note that one of the stated objectives of REF2021 is to "create a strong performance incentive for HEIs and individual researchers"; i.e. the very existence of REF is supposed to improve the performance of HEIs – and not only HEIs but individual researchers too.

Now, academics will know from teaching undergraduates, that grading their work (2:1, 2:2, etc. – i.e. summative evaluation) doesn't actually help them learn, although it might motivate them to improve. What helps us learn is *formative* evaluation – good old-fashioned feedback. And yet it strikes me that very little research evaluation, certainly of the quantitative type, actually does that. It particularly upsets me that the UK spent £250 million on REF 2014 and, for that investment, each institution got a spreadsheet full of numbers and about six lines of written feedback per unit of assessment. This feels like a wasted opportunity.

By contrast, the Dutch Standard Evaluation Protocol (I know, I wax lyrical about this a lot) seems to be a much more formative process. Each research group is provided with written feedback from national and international experts (how many of the latter do we have on our REF panels?). Their past performance and future plans are assessed and those evaluated are given suggestions as to how they might improve. This strikes me as far more collegiate and constructive than rows of grade point averages with no qualitative nuance around strengths and weaknesses, and no real guidance as to how universities might develop.

I would also suggest that written feedback would be more instructive to anyone seeking to understand the strengths and weaknesses of a particular discipline within a university. As it is, if anyone wants to know what Architecture, Built Environment and Planning is like at Loughborough, I will tell you "3.28". And that is clearly all you need to know.

I'm not saying that summative evaluation is never valuable, nor that it's never motivating. (Although I'm yet to find an individual academic who finds the REF motivating in any positive sense.) However, we know from motivational theory that once humans have met certain basic needs (food, shelter, safety) they are almost exclusively motivated by autonomy, connectivity, and competence rather than financial or reputational gain. And I think formative evaluation better enables improved autonomy, connectivity, and competence, than summative evaluation.

I come back to what matters. What matters, I think, is that we serve humanity, that we progress, that we value our researchers and provide an environment in which they can help that happen. And I understand our longing to rank and score. And sometimes we need to do that; there has to be a winner – one job on offer, one grant – and sometimes measuring and scoring can actually make sure we do *equally* value our people. But on the whole, I think formative rather than summative evaluation serves the world better, because it seeks to improve the world rather than judge it.

So, there you have it. My take on how to improve research evaluation in seven hashtags. I'm certain this won't be my final word on the matter and I'm always open to formative evaluation by others. So please feel free to join the conversation and help to co-design a better future for research evaluation!

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

Featured image credit: Jon Tyson, via Unsplash (licensed under a CC0 1.0 license).

About the author

Elizabeth Gadd *is the Research Policy Manager (Publications) at Loughborough University. She has a background in libraries and scholarly communication research. She is the co-founder of the Lis-Bibliometrics Forum and is the ARMA Research Evaluation Special Interest Group Champion.*

Date originally posted: 2018-09-27

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/09/27/better-research-evaluation-in-seven-hashtags/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/