

Although the REF document describes some obvious measurement problems, we are concerned that it does not register the true level of difficulty in designing credible measures for even rough estimates of a feature like 'social, cultural and economic impact' that 1) includes a great variety of very different kinds of effects which, moreover, matter very differently to different aspects of social, cultural and economic welfare and have only at best a very incomplete ordering of importance; 2) includes important very long-term effects; 3) includes contributions that are highly interactive, which makes it especially difficult to estimate what happens with respect to them over longish periods; 4) demands a 'counterfactual' comparison, as impact does, of what is the case and what would have been the case had the research not been conducted.

The broad aim, to take account of the extent to which the HE sector builds on research 'to achieve demonstrable benefits to the wider economy and society', is an abstract one. Designing measures for it requires two steps. First is getting clear what it amounts to in the concrete; second, finding ascertainable measures and indicators that can be shown to be reasonably correlated with the more concretely described aims.

The first step inevitably involves value judgements: What are the kinds of effects that should count as benefits and that we should thus wish to monitor? Often the first step is conflated with the second. Measures and indicators are proposed with only a rough, rather than a fully articulated and agreed upon, account of what the concrete benefits are that we aim to measure. This is dangerous since there is no way to avoid the value implications and consequences of adopting one set of measures over another. Conflating the two steps buries the value judgements implicated in the measures, making them untransparent and removing them from consideration and debate.

The second stage is also delicate, as the REF document acknowledges. We should like to stress the problems involved in finding indicators that can be shown to correlate with counterfactual differences. The measurement aims not to find out if some benefit occurred but rather, as the REF document indicates in discussing *contributions*, to find out what difference the research made to the benefit. To estimate differences like this requires a model of what would have happened without the research in order to compare it with what benefits occurred with the research.

As with the first step, there is a tendency to move too quickly here, to offer measures that seem plausible without producing the requisite models that back them up. But even if we are willing to accept rough measures/indicators, there should be good reason to believe they correlate, at least roughly, with the feature to be measured. Without reasonable, if only rough, models it is difficult to produce credible reasons for thinking any specific indicator correlates with the counterfactual difference we are looking for. The requisite modelling, however, introduces an additional layer of uncertainty/unreliability, particularly when the measures have to take account of a large number of diverse kinds of effects. But winging it without modelling is to bet on what those models must be like and to do so without explicit defence and debate.

Because of these problems we have considerable worry about how accurate results can be. There can be a tendency in the face of these problems to adopt measures with insufficient credibility; for instance, relatively easy-to-access indicators that seem plausible for picking up steps in a possible causal pathway for influence, without modelling how much of the research contribution to all the various different kinds of concrete benefits desired can be expected to flow through these pathways or whether what does flow through these pathways actually contributes improvements.

The amount of collaboration with the kinds of users for whom collaboration makes sense is a possible example here; as is the amount of research funding from users of the type who fund research. In this regard we particularly welcome the REF document's commitment to developing indicators sensitive to contributions that flow through indirect routes.

We are not primarily worried that the results will be imprecise – that they will have wide error bars on each side, but rather that there will be insufficient reason for believing they track the collection of desired benefits even roughly. With insufficient discussion of concrete aims – what is to count concretely as a benefit to be measured – and insufficient reason to think the indicators and measures proposed track these specific benefits, the exercise will inevitably depend heavily on prior views and prior commitments about what research achieves and how; in which case it would be misleading to describe it as a measurement exercise.

Beyond general worries, we should like to raise two fairly obvious issues about specific REF proposals. The first is how to deal with long-term impacts. The document is alert to this but the proposal to consider 'the impact of research undertaken over a sufficiently long time frame' is out of kilter with the rest of the research assessment. Moving back the time frame suggests assessing research outside the period of the rest of the exercise. Moreover, this will very often have been done by people who have changed units since doing the research. Nor does looking at the 'broad impact of the unit as a whole' promise to help much with assessing long-term benefits if the 'broad impact' is impact of the research within the assessment period; if it is earlier research, it suffers from the previous problem. We should like to stress the importance of solving these problems since it seems many of the benefits are, and should be, long term so failing to capture them in a reasonable way could give a much distorted picture of research impact.

Second, the case study method suggested in the REF document is to be applauded but can carry, we worry, a danger of 'undercount'. This method has the great advantage of allowing for diversity of benefit end-points and pathways for contributing to them. But these will have to be constructed for the most part by researchers not expert in thinking about pathways of influence. If panellists have sufficient expertise here they should be good at noting 'overblown' submissions but generally could not be expected to be good at spotting swathes of routes for influences those constructing submissions have not had the ability to recognize. So there is danger of significant 'undercount' even with expert panellists.

Nancy Cartwright FBA
Professor of Philosophy (specializing in methodology, esp. evidence for evidence-based policy)
London School of Economics and University of California at San Diego