The SCOPE Framework

A five-stage process for evaluating research responsibly

scope

Contents

The SCOPE Framework

A five-stage process for evaluating research responsibly

Introduction

The International Network of Research Management Societies (INORMS) Research Evaluation Group (REG) brings together representatives from a range of global member research management societies to work towards better, fairer, and more meaningful research evaluation.

The SCOPE Framework was developed by the REG as a practical way of implementing responsible research evaluation principles to design robust evaluations. We hope this guide will provide a useful steer to research evaluators around the world who are keen to engage with best practice and provide the best service to their organisations.

Dr Elizabeth Gadd,
Chair,
INORMS Research Evaluation Group
August 2021.

Acknowledgements

Why SCOPE?

Research evaluation has historically been indicator- and data-driven and therefore limited to a small number of publication-based citation metrics or research income-based measures. An over-reliance on these indicators has had a significantly negative effect on the research ecosystem: on the types of research, researchers and behaviours, rewarded and embedded. In response to these issues, the community have developed a series of manifestos and principles governing a more responsible approach to researcher evaluation, to the use of metrics in general, and the use of bibliometrics in particular. Many institutions have now adopted a responsible metrics policy based on such principles. However, when it comes to actually applying those principles to the design of alternative research evaluation mechanisms, institutions can often struggle.

As a group of practicing research managers and administrators, the INORMS Research Evaluation Group have sought to resolve this problem by developing a practical five-stage process by which better value-driven research evaluation approaches can be designed. The resulting framework, SCOPE, is an acronym as follows:

- **START** with what you value
- **CONTEXT** considerations
- **OPTIONS** for evaluating
- **PROBE** deeply
- **EVALUATE** your evaluation

How to use this guide

One of the strengths of SCOPE is its simplicity. It's possible to simply keep the five stages of SCOPE in mind as you develop your evaluation: sense-checking whether you are truly measuring what you value, ensuring your evaluation is context-sensitive, considering the validity of the options you use, and double-checking it leads to no unintended consequences. The outside world may not be aware that you are following the SCOPE Framework at all. In this sense, SCOPE acts as the underpinning 'skeleton' for your evaluation design.

On the other hand, you might find it helpful to make your use of the SCOPE Framework visible, perhaps to get buy in from leadership or stakeholders, or to demonstrate that you are approaching your evaluation in a rigorous and robust way. In such cases SCOPE acts as an 'exoskeleton', scaffolding the evaluation from the outside. Those using SCOPE

in a light-touch way, may find the one-page overview of SCOPE a useful checklist of the five stages. However, those wanting to apply SCOPE in a more formal way, may be better supported by the additional detail provided in the full guide.

The guidance provided has been compiled from the results of a series of SCOPE workshops held in different settings with a range of stakeholders: universities, research performing organisations, research funding organisations, and publishers. We're grateful to all those who have been willing to participate in our workshops, resulting in the wisdom now shared with the community.

[1] San Francisco Declaration on Research Assessment. https://sfdora.org/

[2] The Metric Tide Report. https://re.ukri.org/sector-guidance/publications/metric-tide/

[3] The Leiden Manifesto. http://www.leidenmanifesto.org/

A one-page overview of the five-stage SCOPE Framework

START WITH WHAT YOU VALUE

S

EVALUATE YOUR EVALUATION

E

CONTEXT CONSIDERATIONS

C

PROBE DEEPLY

P

OPTIONS FOR EVALUATING

O

The SCOPE Principles

The five stages of SCOPE operate under three main principles:

1. Evaluate only where necessary.
Evaluation is not always the right strategy. When it comes to incentivising behaviours, for example, it may be more fruitful to enable them than to evaluate them.

2. Evaluate with the evaluated.
Any evaluation should be co-designed and co-interpreted by the communities being evaluated.

3. Draw on evaluation expertise.
We should apply the same rigour to our evaluations that we apply to our academic research.

START with what you value

- Clearly articulate what you value about the entity being evaluated
- Not with what others' value (external drivers)
- Not with available data sources (the 'Streetlight Effect')

CONTEXT considerations

- Ensure your evaluation is context-specific
- WHO are you evaluating? (Entity size and discipline)
- WHY are you evaluating?

OPTIONS for evaluating

- Consider both quantitative and qualitative options
- Be careful when using quantities to indicate qualities

PROBE deeply

- WHO might your evaluation approach discriminate against?
- HOW might your evaluation approach be gamed?
- WHAT might the unintended consequences be?
- CONSIDER the cost-benefit of the evaluation

EVALUATE your evaluation

- Did your evaluation achieve its aims?
- Was it formative as well as summative?
- Use SCOPE to evaluate your evaluation.

SCOPE Principles in detail

1. Evaluate only where necessary.

Research and researchers are heavily evaluated: for jobs, grants, promotions, prizes, journal and conference peer review, and management activities. There is a strong sense that over-evaluation is having a negative impact on research culture globally, particularly in terms of mental health. The fact that many of the evaluation approaches used are so poor and narrowly focussed on publication performance, such as the use of journal citation metrics to evaluate individual researchers, only compounds the problem. However, the rush to displace problematic approaches has led to a deluge of new indicators and services evaluating a broader range of things. Some of these may ultimately be helpful, but in the short-term they have led to more evaluation rather than less.

A particularly prevalent use of evaluation is to incentivise the behaviours we would like to see in our research communities. The increase in funder demands around open access is a good example. Whilst this appears logical (Campbell's Law tells us that we get more of what we measure) we should always ensure that we make it as easy as possible for researchers to engage with the desired behaviours before we seek to incentivise them through evaluation. If we think about the take-up of recycling opportunities, most people understood why recycling was important and were keen to do so, but it was only when recycling was made too easy not to do by the provision of household collection services, that most of us started doing it regularly. The best incentive was making it easy and normative, rather than by measuring and rewarding our recycling activity.

For these reasons we encourage those seeking to develop an evaluation using SCOPE to test their thinking by asking 'Do we need to evaluate at all?'.

2. Evaluate with the evaluated.

The benefits of co-design and co-production have been felt in many domains and that of research evaluation is no different. A sector leader here is the Centre for Science and Technologies Studies (CWTS) Leiden and their Evaluative Inquiry approach. To mitigate against concerns around the poor use of metrics and unconscious bias in evaluation, ensuring any approach is both co-produced with, and the results co-interpreted by, the evaluated community is critical.

However, it is important to take a broad view of who the 'evaluated' community is. It has been recognised that our definition of the research community has historically been quite narrow, focussing only on academic researchers, rather than the broader group of technicians, librarians, and research support staff who also make a significant contribution to the research endeavour. There are also other groups that an evaluation might affect, such as data analysts and Human Resources staff, the inclusion of whom in the evaluation design can only improve it.

3. Draw on evaluation expertise.

A criticism often levelled at research evaluation methodologies is that they do not match up to the standards of rigour we would apply to our academic research. Indicators are a poor proxy for the concept they seek to indicate, data is not supplied with error bars, and survey methodologies result in misleading findings. It is important that we are conscious of `epistemic trespass' where expertise in one discipline does not qualify an individual to claim expertise in all, and to draw on the appropriate evaluation expertise when designing our evaluation approaches.

Articulate the evaluation problem

Prior to your evaluation design, you will have some sense as to what it is you seek to evaluate and why. Producing a simple evaluation statement can be a helpful start. So where 'X' is an entity, 'Y' is a purpose, and 'Z' is a behaviour or quality, your statement might read:

We wish to evaluate X for Y.

Example: We wish to evaluate researchers for recruitment.

Or

We wish to Y Z.

Example: We wish to monitor open research.

Writing it out like this can help you to start to define your evaluation before you launch into SCOPE. For example, if you seek to evaluate researchers for recruitment, you can define what career stages you're looking at, or whether the criteria might be useful beyond recruitment, perhaps for appraisal or promotion. Similarly, by articulating the evaluation problem, you might discover that the problem is not an evaluation problem at all, but a management problem or a practical one. For example, at the outset you may seek to monitor open research but soon realise that it's actually engagement with open research that you want, and evaluation is not the best way of achieving that.

Taking a SCOPE workshop approach

Taking a workshop approach to your evaluation design is a good way of 'evaluating with the evaluated'. Workshops provide a way of ensuring both the evaluated, and the relevant stakeholder communities, have a say both as to what they value and how they might evaluate it. The discussions generated can build confidence in the process, and consensus as to the best way forward. However, there is a limit on the number of people who can attend a workshop, so some organisations choose to supplement their workshops with online surveys. Surveys can provide everyone with an opportunity to input into the evaluation design. They are also less subject to the dangers of 'groupthink'.

There is no one single approach to running SCOPE workshops but a model that has worked well for us is as follows:

A core evaluation team is established at the evaluating organization.

The core team identify a diverse cross-section of the evaluated community for the first workshop.

Run Workshop 1: What do you value about the entity you seek to evaluate?

The core team synthesise the findings from the workshop and share the resulting list of values with participants for comment.

Run Workshop 2: Identify options for evaluating and probe those options.

The core team identify a suitable group of both evaluated and other stakeholders for a second workshop.

Once the values are agreed, core team identify and/or confirm the contexts in which they wish to evaluate.

The core team synthesise the outcomes of the workshop and design an evaluation approach.

The core team share the evaluation approach with all participants from Workshops 1 and 2 for feedback.

Implement the evaluation approach.

The core team evaluate the evaluation and share the findings with the evaluated and stakeholder communities.

A few things to bear in mind:

- Larger evaluation exercises may need more than two workshops to design.

- Whilst seeking to evaluate with the evaluated, you will still have evaluation goals to meet. Stakeholders should inform but not dictate your approach.

- Evaluation discussions can often raise a lot of related issues that are not central to the evaluation design. Be prepared to 'park' these for consideration in another forum.

Start with what you value

The rationale:

The reason so many forms of research evaluation have a negative impact on research culture is that they do not start with what is actually valued about the entity under evaluation. Instead, evaluators may start with the data sources they have available to them – often bibliometric data – or with the values of third parties such as government funders and university ranking agencies. Whilst such influences cannot be ignored, they should not form the starting point of any evaluation. For this reason, the S of SCOPE is to 'Start with what you value' about the entity under evaluation. The other important consideration here is to keep those values under review. What an evaluator valued even three or five years ago may not be what is valued now.

Understanding what we value

The first step in understanding what we value about the entity under evaluation is to consider, who's 'we'? For example, a research funder may be the body undertaking the evaluation, but they may be funding research that seeks to deliver benefits to recipient communities. Both stakeholder groups will have a perspective as to what research should be funded. Within a recruiting organisation, both the senior leadership, the disciplinary leads and HR colleagues will have a perspective as to what the organisation should be looking for in new research staff. The approach we usually take with SCOPE evaluations is to involve all stakeholders in developing an agreed sense as to what is valued about the entity under evaluation. In line with our first principle of evaluating with the evaluated, we recommend this process includes representatives of the entity under evaluation.

What is a value?

Simply put, a value is a judgment made about what is important. However, value judgments can be made at different levels of granularity. In our work we've observed three different layers of values, all of which are helpful to the evaluation process.

SUPER-VALUES – At the highest level, we have what might be called 'super-values'. Most organisations have thought about their values and may have a short list of them, often stated as single words, such as collegiality, transparency, or inclusivity. These can be useful in steering the evaluation but don't usually provide enough detail to design an evaluation.

VALUES – At the next level down we have our values themselves. We'd describe these as the things you care about, given your super-values, and the things you want to evaluate. You can understand your values by asking how your super-value(s) manifest themselves.

SUBVALUES – At the lowest level of granularity, values might be seen to be divided up into sub-values. Understanding what your values 'look and feel like' in your organisation will provide an in-depth understanding of your values which can be helpful in an evaluation.

Exploring super-values, values and sub-values at the University of Glasgow

The University of Glasgow's research strategy seeks to embody three key 'super-values': Collaboration, Creativity and Careers. In an effort to embed these super-values in their institution they identified an ambition to encourage 'support for the careers of others'. This was the value they sought to evaluate; the embodiment of their super-value. Having identified this value, they sought to understand its dimensions in the Glasgow context through a workshop with stakeholders. The final list of sub-values were as follows:

1. The act of giving something up or going out of your way to benefit someone else's career (e.g., time to mentor, an opportunity, authorship on publication list), with the aim of increasing one's "net" contribution to the research system.

2. Taking a coaching approach to career support, i.e., setting the expectation that a mentee will be allowed to, and be supported to, explore opportunities relevant to career development and that a mentor need not have specific knowledge to support that development.

3. The act of lobbying externally to enable better support for researcher careers (either within the HEI e.g., creating more permanent posts, or externally e.g. addressing the steep career pyramid).

4. Thinking about & valuing careers both within and beyond academia.

Exploring what you value

Understanding what you value about the entity you're seeking to evaluate is probably the most challenging step in any evaluation design. This is probably why the process is engaged in so rarely. However, we'd encourage evaluators to give this some serious thought at the outset of the evaluation process and not rush on to the next step too soon. Some of the questions we have used in workshops to get to the heart of what we value about a particular entity are as follows. You might want to insert the words 'open research' or 'research culture' to see how these questions might be applicable.

- **What value results from [X]?**
- **What are the negative outcomes of a poor [X]?**
- **What does a positive [X] look like?**
- **What does a negative [X] look like?**

- **How can you tell when an organisation has a positive [X]?**
- **How can you tell when an organisation has a poor [X]?**
- **Why would we care about [X]?**

When thinking about what is valued about a particular entity, it can be helpful to also explore what is not valued, and how the resulting values might be weighted. For example, an organisation might value research leadership but if they also value publication performance, they may need to decide which has the greater weight in an evaluation situation: fewer primary authored publications because the researcher has given primary authorship to their research assistants, or a long list of primary-authored publications? To identify how stakeholders might weight the things they value, this can be built into the values exercise by means of voting or identifying the frequency of recurring terms.

Case Study

Understanding the 'look and feel' of research culture values at Newcastle University

Newcastle University used SCOPE to explore how they might evaluate their progress towards building a more positive research culture. To this end, they held a workshop with a range of colleagues, from both academic and professional services communities, and identified four key 'super-values':

- **Collaboration & collegiality**
- **The freedom to explore & grow**
- **Fairness & inclusion**
- **Openness & integrity**

To thoroughly understand the dimensions of these super-values in the Newcastle context, they also explored what each might 'look like' and might 'feel like'. For example, colleagues felt that 'Openness and integrity' might involve 'transparent internal processes and decision-making' and feel 'rigorous, non-judgmental and safe'. By exploring both dimensions of their super-values – the behaviours and structures that may result, and the emotional and psychological impacts – they were in a good position to ensure that any resulting Options for evaluating supported both these things. This work will underpin an institutional 'Research Culture Index' to monitor research culture change across a basket of values-led measures.

Context considerations

The rationale

It's common to witness arguments as to which metrics are 'responsible' and which metrics are not, without any reference being made to the contexts in which those metrics are being used. A good case in point is the use of Elsevier's Field-Weighted Citation Impact (FWCI) indicator. This is only really suitable for assessing the relative citedness of large publication sets. As such, when used to compare the citedness of one country with another, it can be a useful indicator. However, when it is used to identify individual researchers for promotion or redundancy it is deeply problematic. Evaluation approaches have to be context-specific across two main dimensions: who are you evaluating and why?

Why are you evaluating?

It is generally accepted that there are six key reasons why you might seek to evaluate .

- **ANALYSIS**. Evaluating to understand. "Science of science" activities that study patterns and trends for the sole purpose of understanding them better.

- **ADVOCACY**. Evaluating to 'show off'. Evaluation activities that seek to highlight an entity's strengths for cases such as promotional materials or grant applications.

- **ACCOUNTABILITY.** Evaluating to monitor. Plotting progress against an objective whether internally or externally set.

- **ACCLAIM.** Evaluating to compare. Assessments that seek to compare or benchmark one entity with another. University rankings are an example of this.

- **ADAPTATION.** Evaluating to incentivise. The use of assessment to incentivise certain behaviours. Funder assessments of engagement with open access is one example of this.

- **ALLOCATION.** Evaluating to reward. Any activity that results in some form of reward for the entities being evaluated, be this a job, promotion, grant, prize or award of any description.

Case Study

Evaluating to incentivise at Emerald Publishing

Emerald Publishing sought to incentivise more diverse editorial boards on their journals as part of a larger commitment to equality and diversity. By running a SCOPE workshop with editors and editorial board members it was recognised that whilst evaluation may play a part in incentivising more diverse editorial boards there were also a whole host of enabling actions that needed to take place at the same time. This chimes with the first principle of SCOPE to evaluate only where necessary. They also learned that with a long-term goal such as enabling diversity, evaluating a commitment to that goal rather than an attainment of that goal is a useful first step.

Who (or what) are you evaluating?

There are two key considerations when defining the entity being evaluated. The first is size and the second is discipline.

Clearly when it comes to research indicators, the smaller the 'sample' or entity size, the less confident the evaluator can be that indicators can tell us something meaningful about that entity, rather than being down to chance. For example, the number of publications produced by two researchers of a similar career stage and discipline could vary significantly due to a large number of extraneous variables (illness, opportunity, family/career leave), However, the number of publications produced by two universities are less likely to be negatively affected by such factors as the two samples are likely to be subject to the same variables overall.

The second consideration when thinking about context - and something explicitly picked up by the Leiden Manifesto - is that of the entity's discipline. Different disciplines operate under different research paradigms, have different funding opportunities, use different methods, and take very different approaches to publication. An evaluation approach that is entirely sensible in one discipline could be completely inappropriate to another. This is another occasion on which our principle of evaluating with the evaluated ensures that such differences are surfaced and understood. Similarly, the principle of drawing on evaluation expertise can ensure that the implication of such differences for evaluation purposes can be considered.

The evaluation impact matrix

The reason it's important to identify who (or what) and why you are evaluating prior to considering how you might undertake the evaluation is that assessments in some settings have more impact on the entity being evaluated and are therefore a higher risk. To provide a sense as to where these boundaries lie, and where there may be greater impacts and risks, we have plotted the six different evaluation purposes against four different entity sizes on the following matrix. Each segment has been 'RAG' (red-amber-green) rated to indicate whether an evaluation in that context might be high, medium, or low impact. The RAG-rating is provided as a guide as the continuum of impact/risk across various evaluative settings rather than a rule.

An entity's discipline should be borne in mind across all these segments and, depending on the values being evaluated, will affect the options chosen in the next stage.

An evaluator should identify the contexts in which the evaluation will take place prior to moving on to the Options for evaluating. If seeking to evaluate in multiple contexts, you will need to think separately about the options for evaluating in each.

		Country	HEI	Group	Individual
Analysis	To understand	Low	Low	Medium	Medium
Advocacy	To show off	Low	Low	Medium	Medium
Accountability	To monitor	Low	Medium	Medium	High
Acclaim	To benchmark	Medium	High	High	High
Adaptation	To incentivise	Medium	High	High	High
Allocation	To reward	High	High	High	High

- 🟩 Low impact
- 🟧 Medium impact
- 🟥 High impact
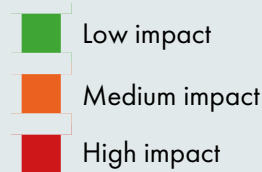
Options for evaluating

The rationale

The purpose of the Options stage is to explore all the options – both qualitative and quantitative - for evaluating your identified values in the chosen contexts. As a 'rule of thumb' it is suggested that quantitative indicators are better for quantitative things: citations, research income, student numbers, and qualitative approaches are better for qualitative things such as research impact. Evaluators should exercise particular caution when using quantitative indicators for qualitative things. For example, citation counts are not a proxy for quality, and a university's ranking position does not indicate excellence.

Exploring your options

Given the many and varied values and contexts which may be evaluated, it is not possible to provide a comprehensive list of options for doing so in this guide. However, there are sources of alternative evaluation approaches such as those provided by the DORA resource library and the Metrics Toolkit which can offer some inspiration.

At this stage it may be helpful to generate a number of different options given that the Probe stage will stress-test these options and may render some unusable. The alternative is to consider both the Options and Probe stages together. This is the model we use for SCOPE workshops, ensuring that no option is developed so far that it becomes too difficult to abandon it after being 'probed'.

It is worth remembering that no one indicator or evaluative approach is going to represent everything you value about the entity being evaluated. What is being sought is a proportionate and appropriate assessment that will (as with the Hippocratic Oath) first do no harm. This will always involve human judgement in some form, and always involve an approximation of the reliability of the assessment (error bars, list of caveats/limitations, etc.,).

One of the common Options considerations when seeking to deliver a value-led evaluation is whether to assess evidence of a particular value or commitment to that value. This is particularly important when evaluating early career researchers that may not yet have evidence to present but have the potential and the commitment to a particular value. It's also pertinent when seeking to incentivise behaviours where there may not yet be equal opportunity for all groups to engage with the new activity.

Other Options considerations to bear in mind in different contexts are as follows:

• ANALYSIS
Evaluating to understand.

Evaluations of this nature are likely to require some form of data analysis. It's worth remembering that you may be able to learn from published studies in a similar area rather than having to gather more primary data yourself.

• ADVOCACY
Evaluating to 'show off'.

Evaluations that seek to generate evidence for advocacy purposes (promotional activities etc.,) often need only to highlight the entity's strengths rather than present an holistic analysis of their performance. For example, they may point to external prizes or impressive numbers (e.g., 90% of our outputs are available on open access).

• ACCOUNTABILITY
Evaluating to monitor.

Evaluations that seek to monitor progress against a particular goal require reliable data that will be available over a period of time for trend analysis. This should be factored into any option chosen. In some cases, targets may be set (e.g., KPIs). Whilst this can provide the motivation to improve on a particular objective, targets can be problematic if cascaded down to individual researchers. Thus, a goal of improving a department's average research grant income per FTE might be a legitimate one, but it should never be translated as a need for all researchers to generate a target level of income. There will be legitimate reasons for variability amongst individual researchers and research groups.

• ACCLAIM
Evaluating to compare.

Comparative evaluations require reliable sector-wide data or intelligence. Such evaluations are best used to identify strengths in an entity and to learn from their best practice rather than as a punitive exercise that highlights where entities do not 'measure up'. Benchmarking at the level of individual researchers should generally be avoided unless there is a need to select a 'winner' (see Allocation below). As a general rule, if the evidence for an entity looks positive (e.g., their citation performance is strong), it usually means something (although what that 'something' is should be carefully determined); if the evidence does not look positive, it probably means nothing (e.g., they are publishing in citation scarce fields).

• ADAPTATION
Evaluating to incentivise.

In one sense all evaluations form an incentive on some level as all evaluations trickle down to individual researchers. However, as mentioned above, it is important to ensure that any hoped-for behaviour is first enabled before it is evaluated. Evaluative incentives might include prizes (e.g., research culture awards), required thresholds (e.g., Funder open access policies), or audits and need to be appropriate to the importance of the incentivised behaviour.

• ALLOCATION
Evaluating to reward.

Selecting candidates for jobs, prizes or grant income usually requires thresholds (who makes the shortlist) and ranking (who gets the reward). Indeed, it is the only activity on the evaluation spectrum that necessitates ranking. However, due to the increasing competitiveness in academia, some funders are now exploring the use of lotteries to allocate grants to those that meet the required threshold.

Working with data and indicators

Often our value-led evaluations will lead us to evaluate new dimensions for which there are no readily available sources of data. If you find yourself in this situation you might want to pursue one of the following options:

- **LONG TERM.**
  You could seek to collect the data either through an annual survey or by partnering with a third party.

- **MEDIUM TERM.**
  You could run a one-off survey or data collection exercise to generate some snapshot data.

- **SHORT TERM.**
  You could see if there are any existing secondary data sources, such as those accompanying published studies, which could inform your evaluation.

When using quantitative approaches it is important to keep responsible metrics principles in mind. In particular, the fact that the smaller the entity/sample size, the less suitable quantitative approaches (such as bibliometrics) are. If using data for small sample sizes/entities is unavoidable they should be shared with the evaluated for checking, and they should have an opportunity to provide free-text comments. A useful checklist when designing indicators is as follows:
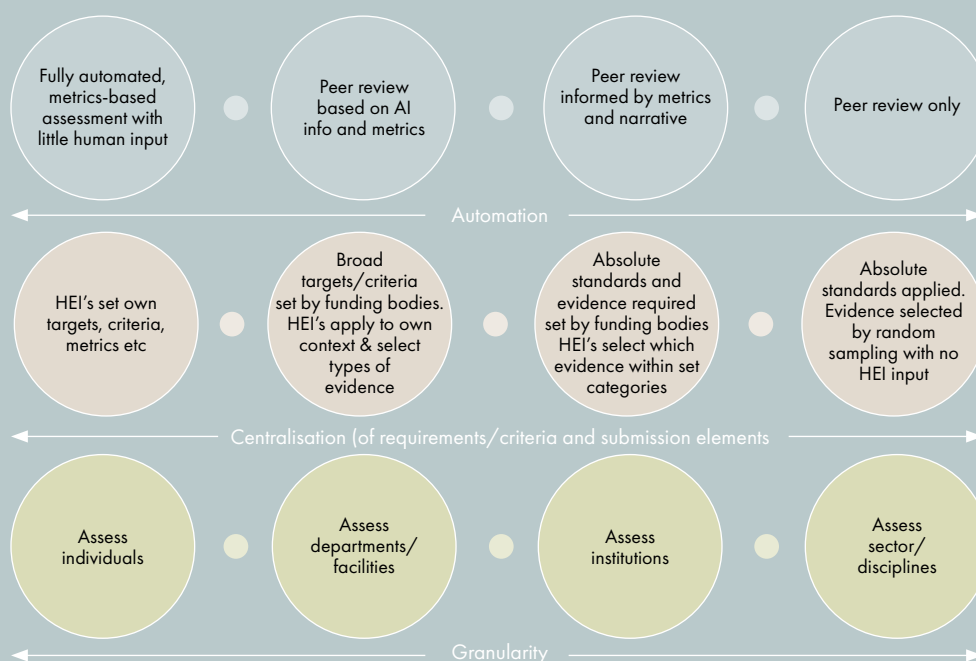
Indicators should be:

- **VALID** (reflecting the concept measured)
- **UNDERSTANDABLE**
- **TRANSPARENT** (data underlying criteria should be released, with clearly explained limitations and degrees of uncertainty)
- **FAIR** (systematic bias should be avoided)
- **ADAPTIVE** (updated when bias, abuse or other weaknesses become apparent);
- **REPRODUCIBLE** (those who use the indicator should be able to reproduce it).

Taken from: Wouters et al, 2019. Rethinking impact factors. Nature. 569, 621-623 (2019) doi: https://doi.org/10.1038/d41586-019-01643-3

Case Study

The four UK research funding bodies – Assessing stakeholders' appetite for various evaluation options

The four UK research funding bodies used SCOPE to underpin their Future Research Assessment Programme (FRAP). A series of roundtables were held with a range of stakeholder groups at which they sought to identify the appetite for different variables that might impact any future Options for evaluating. The spectra addressed elements such as frequency, automation, centralisation and granularity. This was an excellent way of getting early input on evaluation options prior to the actual design of the approach.

Fully automated, metrics-based assessment with little human input	Peer review based on AI info and metrics	Peer review informed by metrics and narrative	Peer review only

Automation ←———————————————————————————→

HEI's set own targets, criteria, metrics etc	Broad targets/criteria set by funding bodies. HEI's apply to own context & select types of evidence	Absolute standards and evidence required set by funding bodies HEI's select which evidence within set categories	Absolute standards applied. Evidence selected by random sampling with no HEI input

Centralisation (of requirements/criteria and submission elements) ←———————————————————————————→

Assess individuals	Assess departments/facilities	Assess institutions	Assess sector/disciplines

Granularity ←———————————————————————————→

Spectra used by the four UK research funding bodies FRAP Roundtables to assess appetite for various evaluation options

Probe

The rationale

Many of the problematic approaches to research evaluation that currently dominate the research ecosystem could have been avoided if they were 'probed' for harmful impacts and possible unintended consequences at their inception. The four questions that we think are key to ask of any options for evaluating are as follows:

1. Who might this discriminate against?

When instituting new systems and procedures it is good practice to run an Equality Impact Assessment (EIA) on them. EIAs seek to identify where a process or activity has unequal impacts on different groups of people thereby unintentionally building in inequality. This question acts as an EIA on the evaluation design, seeking to ensure it does not unintentionally discriminate against particular disciplines or demographics.

2. How might this be gamed?

One of the biggest complaints about indicators is that they frequently lead to gaming behaviours. Campbell's Law tells us that we get what we measure and evaluation approaches that seek adaptation rely on this. However, Goodhart's Law reminds us that when a measure becomes a target it ceases to be a good measure. We should therefore always seek to design out gaming wherever possible.

3. What might the unintended consequences be?

The unintended consequences of evaluation approaches are, by definition, unexpected and often unpredictable. However, in many cases some forethought and stress-testing can help us to identify where an evaluation approach might have negative long-term consequences.

4. What is the cost-benefit?

There is no doubt that the more precise, more qualitative approaches to research evaluation are more expensive than data-driven approaches based on existing (often bibliometric) sources. There is also a law of diminishing returns when introducing new measures: whilst they might identify some areas for immediate improvement, once those areas have been addressed there may be little point continuing with the evaluation in its current form. No evaluation is perfect and at some point an assessment needs to be made as to what is proportionate given the potential consequences of the evaluation.

Case Study

Evaluating open research at Loughborough University

Loughborough University has an Open Research Position Statement and used the SCOPE framework to identify suitable ways to evaluate progress against their open research ambitions. Having considered some Options for evaluating, they Probed those options using the four questions and identified the following potential issues:

1. Who does this DISCRIMINATE against?

Not all disciplines have an equal opportunity to engage with Open Research.

2. How might this be GAMED?

Assessing the percentage of an individual's open access outputs on the University CRIS system may lead to closed outputs being left off the CRIS.

3. What might the UNINTENDED CONSEQUENCES be?

There is a danger of creating an 'and' culture where openness is seen as an additional box to tick rather than an alternative way of doing research.

4. What is the COST-BENEFIT?

There are currently very few reliable data sources to fairly assess engagement with open research practices. Much would therefore rely on costly, manual data collection.

As a consequence of the Probe stage and given an ambition of seeking 'Recognition over Judgement', an initial approach was agreed on to celebrate Open Research success where it was found by instituting some Open Research awards.

Probing your evaluation

1. Assessing for discriminatory effects

The discriminatory effects of some forms of evaluation, such as bibliometric approaches and peer review, are well-documented. Such evidence can support evaluators to put in place mitigating actions such as double-blind peer review, EDI observers, and guidance for assessors. For new forms of evaluation, it can be helpful if a small pilot evaluation is run and the outcomes assessed to see what the impacts might be. If a pilot is not an option, it is sometimes possible to predict unequal outcomes by running a thought experiment such as the ones described below.

2. Assessing for gaming potential

One way to design out gaming in an evaluation is to regularly change the evaluation approach. Obviously, this is not helpful if you want to build up data over time to monitor trends. However, it could ensure that entities don't spend time pursuing activities that lead only to improvements on key indicators and not to their research practice.

3. Assessing for unintended consequences

One tool that can be useful in identifying unintended consequences of an evaluation design is the following use/misuse/abuse/stress case matrix developed by Eric Meyer and Sarah Wachter-Boechtter[10].

Thinking about how an evaluation approach might be deliberately abused, unintentionally misused, or used in the right way for the wrong purpose can help you to predict and therefore mitigate against some unintended consequences.

Another way of approaching unintended consequences is to explicitly explore the following common unintended consequences of research evaluation adapted from Jerry Z. Muller[11].

- **Goal displacement** – what are you not evaluating that may get overlooked as a consequence?
- **Short-termism** – what long term aims may be missed as a consequence of focussing on short-term evaluation goals?
- **Discouraging risk-taking and innovation** – will the evaluation work against creativity and serendipitous opportunity-taking?
- **Discouraging co-operation and common purpose** – will the evaluation lead to greater co-operation or less?

4. Assessing cost-benefit

Identifying the cost-benefit ratio of any evaluation is a challenging but important exercise. Costs will vary according to whether using existing or new data sources and increase considerably if there are any peer review elements. The benefits should take into account the impact of the exercise on the evaluated entities (such as funds being distributed, the significance of any prizes) and the evaluating organisation (the significance of an appointment and the long-term investment being made). To get the most benefit from an evaluation, it is sensible (where possible) to design in a formative element rather than it being a simply summative exercise. Evaluations that provide the evaluated with guidance as to how to improve can lead to greater long-term benefits than those that simply judge historic performance.

[10] Meyer, E. & S. Wachter-Boettcher. (2016) Design for Real Life. A Book Apart.

[11] Muller, J. T. (2018). The Tyranny of Metrics. Princeton Press.

How we want it used

Stress Case	Use Case
Misuse Case	Abuse Case

Unintentional — Deliberate

How we don't want it used

Adapted by Dawn Duhaney, based on Meyer & Wachter-Boechtter (2016) Design for Real Life.

Evaluate

Having designed your evaluation, the final stage of SCOPE is to run the evaluation, and then evaluate your evaluation. This is particularly important from a Probe perspective as often unintended consequences do not come to light until after the evaluation has been performed. SCOPE can be used just as successfully to evaluate an evaluation as well as to design one. With this in mind you may wish to ask:

- **What VALUE** did you get from the evaluation? Did it generate useful intelligence/outcomes in line with the values you sought to evaluate? What did you anticipate that success would look like?

- **In what CONTEXTS** might you evaluate your evaluation? At what level of granularity (researcher, group, department) and for what purpose (advocating for the initiative or monitoring progress).

- **What are your OPTIONS** for evaluating your evaluation? When might be a suitable time to run the evaluation? Are there existing mechanisms you can tap into, such as staff surveys, or data sources, to assess how successful the evaluation has been?

- **Can you PROBE** the evaluation outcomes to identify any unintended consequences or discriminatory effects?

Having evaluated your evaluation, you are then in a position to either redesign it or rerun it.

Case Study

Evaluating University of Glasgow's evaluation of support for the careers for others

After identifying a mechanism for evaluating researchers' support for the careers of others, Glasgow's attention turned to the process by which they might evaluate their evaluation.

Some of the questions they asked themselves were as follows:

- **START with what you value. Was it:**
  - > Individuals committed to supporting the careers of others?
  - > Careers that have actually been supported?
  - > Creating an expectation, a 'culture', of support?
- **CONTEXT considerations**
  - > At what level of granularity did they wish to evaluate this evaluation?
    - o Individuals? Departments? University wide?
  - > For what purpose?
    - o Success stories? Monitoring progress?
- **OPTIONS for evaluating the evaluation**
  - > Dedicated or existing staff surveys? Case studies? Faculty reviews?
  - > Through new or existing data? Destinations of graduates/staff; Increased take-up of coaching.
- **PROBE deeply**
  - > Will this evaluation approach provide the required answers at a 'price' they are willing to pay?