



## UvA-DARE (Digital Academic Repository)

### Benchmarking in research evaluations

*we can do without it*

van Vree, F.; van Leeuwen, T.

#### DOI

[10.59350/h4f1k-00y10](https://doi.org/10.59350/h4f1k-00y10)

[10.59350/crbvf-qkw52](https://doi.org/10.59350/crbvf-qkw52)

#### Publication date

2025

#### Document Version

Final published version

#### License

CC BY

[Link to publication](#)

#### Citation for published version (APA):

van Vree, F., & van Leeuwen, T. (2025). Benchmarking in research evaluations: we can do without it. Web publication or website, Centre for Science and Technology Studies (CWTS). <https://doi.org/10.59350/h4f1k-00y10>, <https://doi.org/10.59350/crbvf-qkw52>

#### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

#### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

# Benchmarking in research evaluations: we can do without it

**The van Leeuwen** and **Frank van Vree**

Published January 30, 2025

## **Citation**

van Leeuwen, T., & van Vree, F. (2025, January 30). Benchmarking in research evaluations: we can do without it. *Leiden Madtrics*. <https://doi.org/10.59350/h4f1k-00y10>

## **Copyright**

Copyright © The van Leeuwen et al. 2025. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## Leiden Madtrics

[\(Klik hier om de Nederlandse versie te lezen\)](#)

In any evaluation of scientific research, the obvious question is: how does a research unit compare with others, nationally and internationally? This question is easier to ask than to answer, because the answer is not easy to substantiate. The [Strategy Evaluation Protocol \(SEP\) 2021-2027](#), which provides guidelines for the mandatory periodic evaluation of the research of Dutch universities and research institutes of the KNAW and the national research fund NWO, refers to *benchmarking* as a possible way of generating robust data, but does not explain what exactly is meant by this technique/methodology and how it can be operationalised. This is not without danger, as there are a number of objections to *benchmarking* in the strict sense of *comparing performance on the basis of quantitative criteria (time, cost, revenue)*. In view of the ongoing preparation of the new SEP for the period 2027-2033, it is therefore advisable to critically review benchmarking as a method, without wanting to completely ban the idea of comparison as such.

### A problematic tool

On page 19 of the SEP, *benchmarking* is introduced as a possible way of generating robust data: 'Other sources of robust data may include benchmarking against peer research units [...]'. In other words, in addition to the indicators and case studies selected by the unit to demonstrate the quality and impact of the research, benchmarking can be used to support the narrative arguments in the self-evaluation. However, what exactly is meant by benchmarking is not made clear, but it is suggested that benchmarking and quantitative indicators, in a kind of hybrid form, can provide effective support.

This assumption is problematic, to say the least. It starts with the fact that it proposes a technique without specifying how it should be operationalised. This problem is exacerbated by the common definition of benchmarking: the comparison of an institution's or company's business processes and performance statistics with the *best practices* of other companies and institutions in the same or a similar industry, usually measured in terms of quality, time and cost. However, the purely quantitative criteria on which such performance comparisons are usually based are at odds with the spirit and purpose of the SEP. Indeed, the emphasis is on assessing quality and impact from the perspective of one's own mission and related strategy, for which the use of seemingly 'objective' metric criteria (JIF, [h-index](#)) is considered inappropriate.

But this is certainly not the only problem. The lack of a clear definition and operationalisation suggests that such an operation based on quantitative data is relatively easy, when in fact it is extremely problematic and therefore inadvisable on both substantive and ethical grounds.

### Unequal data

The lack of a clear definition and operationalisation of benchmarking in the context of research evaluation means that there are no criteria to determine what the benchmarking material

## Leiden Madtrics

should meet. This opens the door to randomness, arbitrariness and opportunism. This applies first of all to the choice of research units to be included in the benchmarking. An unambitious choice of benchmark units may produce a good result in the evaluation process, but not necessarily the best result in the longer term; conversely, an overambitious choice may lead to a negative result where it should not have been necessary.

At least as problematic in quantitative benchmarking is the disparity in the availability and nature of the underlying data. With regard to one's own research unit, one often has a good insight into the collected material, which comes from local information systems (such as [Pure](#) or [Metis](#), but also [Converis](#)) and is compiled according to known criteria, but this information is rarely available when it comes to the data of the units used for comparison in the benchmark. This is all the more true for institutions abroad, where other systems and standards prevail. The lack of understanding of how the data have been collected and what value they represent seriously compromises the validity of the final comparison, often without people being aware of it.

A comparison based on quantitative indicators assumes that these indicators represent the units in a similar way. However, this is only true to a very limited extent. This problem is well illustrated by the practice of field normalisation, which is used in bibliometric studies to compensate for differences in referencing behaviour in different publication and reference cultures by specialty. As long as the comparison is between two related fields, such as cardiology and oncology, such normalisation makes sense for comparison purposes. However, the situation is different when a study analyses and compares different disciplines - or even whole universities. Theoretically, one could compare citation numbers after normalising for field, but this ignores the sometimes huge differences in publication culture. For physicists, for example, you might find that 80-85% of all publications are indeed published in international journals, which gives a reasonably good picture of the output of the field. However, for many fields within the humanities, such as history and literature, the proportions are quite different, with a significant proportion of output being published in the form of books and book chapters, often in languages other than English, which do not appear in systems such as Scopus and Web of Science. Such differences even exist within disciplines, with more application-oriented specialisms using very different communication channels. In short, making quantitative data 'comparable' is likely to give a biased and distorted picture of the quality and impact of research units and the strategic choices they advocate.

## Ethical concerns

A final point concerns the ethics of this process. The data from one's own unit have, if all goes well, been checked and validated as valid material for evaluation purposes. This is not the case for the material used for the benchmark units: it may have been collected in a different way or for a completely different purpose. Quality checks on this material would therefore be a prerequisite, but this obviously requires the consent of the owner of the data. Apart from that, the use of data without consent is unethical, especially in view of the fact that in the Netherlands the results of the evaluation have to be made public, which could damage the reputation and image of the benchmark units.

### Conclusion

From the above, we can only conclude that benchmarking in the sense of *comparing performance on the basis of quantitative criteria (time, cost, return)* is in most cases inappropriate for evaluations in the sense of the SEP, both for theoretical and ethical reasons. This is not to say that comparison or, if you like, *benchmarking* in a less strict sense of the word, cannot be useful. One's own position, mission and strategy should serve as a starting point, as a way of sharpening the research unit from a more qualitative perspective. Such a mirror, for example in the form of a case study, can be extremely useful.

Header image: Gowtham AGM on [Unsplash](#).

DOI: 10.59350/h4f1k-00y10 ([export/download/cite this blog post](#))