https://doi.org/10.51408/issi2025_153

Responsible Research Assessment of Teams: Reflections and Perspectives After Two Evaluation Cycles at the University of Antwerp, Belgium

Tim C.E. Engels[1], Birgit Houben[2], Pieter Spooren[3]

[1]*tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, and Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

[2]*birgit.houben@uantwerpen.be*
Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

[3]*pieter.spooren@uantwerpen.be*
Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

The University of Antwerp started research assessments per discipline that include a site visit by an international panel of peers in 2007. A few years later we reported that for research teams in the sciences basic metrics like group size, h-index and efficiency in publishing in top journals predicted panel assessments of quality and productivity (Engels et al, 2013). Upon the completion of the second cycle of research assessments in the current academic year 2024-2025, we ask ourselves (1) to what extent the stated aim of improving the quality and impact of research has been achieved, and (2) what shape the third cycle of research assessments will take. For this third cycle, the need to reconcile quantitative and qualitative approaches, responsible use of metrics, transparency and inclusivity are top priorities.

In this paper we first analyze and reflect upon the evolution, the results and the lived experiences of the UAntwerp research assessments since 2007. We then present our proposal for the third cycle of UAntwerp research assessments that will focus on creating contexts in which research can flourish. To achieve this, the assessments will take achievements and bibliometric and other indicators as context elements rather than as elements of assessment. Our aim is to launch a system of more responsible research assessments that will be fully formative and future-oriented, with validated dashboards capturing inputs, process elements, outputs, and impacts as context elements for qualitative assessments.

Introduction

Research assessment needs to consider input, process, output and impact of research, whereby impact involves both scholarly-academic impact as well as broader cultural, economic and social impact (Moed, 2017). In practice, however, research assessment too often mainly relates to bibliometric indicators of journal articles indexed in citation indexes such as Web of Science or Scopus. Even though bibliometricians have repeatedly stressed the important limitations of the use of bibliometrics when assessing individual researchers (e.g. Wouters et al., 2013), it seems that the omnipresence of bibliometric indicators has taken over research assessment at many

levels (Wilsdon et al., 2015), leading to calls to seriously rethink their use (CoARA, 2022; Zhang & Sivertsen, 2020), as well as fierce debate about bibliometrics versus peer review (Abramo, 2024).

In our modest opinion, a debate that is just as important is how to conceive responsible research assessments that do consider input, process, output and impact of research, thereby integrating qualitative and quantitative approaches. Since research assessment and research behaviour co-evolve (OECD Global Science Forum, 2025), and science has become a team effort in a majority of cases, there is a pressing need to rethink research assessment of teams. Such research assessments should welcome inclusive research and a diversity of outputs and impacts, while emphasizing the importance of a research context, research environment, and research process conducive to responsible research and innovation with impact. In this paper we explore, after two cycles of research assessments of teams at the University of Antwerp, how the next cycle of research assessments at our university can be brought more in line with the ambitions of the responsible research assessment agenda (Global Research Council, 2024).

Investments in and evaluation of university research in Flanders

According to the regional innovation scoreboard of the European Commission, Flanders is an innovation leader. The region scores particularly well in terms of international scientific co-publications and public-private co-publications, illustrating the large extent of internationalization of university research and the strong integration of innovation ecosystems in the region. Although government expenditures on R&D remain well below 1% of the regional GDP, business expenditures on R&D have increased significantly over the last decade and are currently well above 3% (IDEA Consult, 2024) As for Europe as a whole, boosting productivity and competitiveness are major challenges, all the more so since increased private investments in R&D seem not to translate in productivity increases as expected.

Flanders is well known for its system of performance-based funding of university research (Debackere & Glänzel, 2004; Engels & Guns, 2018). At the occasion of the Nordic Workshop on Bibliometrics and Research Policy 2023, Engels & Guns analyzed the co-evolution of the PRFS with bibliometric performance indicators and reported an initial increase in per capita productivity. In the longer term, scholarly productivity and impact seem to have stabilized at a relatively high level, which also shows in the bibliometric indicators of the aforementioned regional innovation scoreboard. Holding such a position becomes less evident given the intense global competition for talent and infrastructure in science and technology, and may over time result in a slight or gradual reduction of competitiveness.

Less well known than the Flemish PRFS is that universities in Flanders have a legal obligation to assess the quality of their research activities (Engels et al., 2013). These research assessments resemble systems in the Netherlands and Norway (Sivertsen, 2017), whereby research assessment at the level of departments or research teams takes place without direct financial consequences for the university or the departments and teams involved. In other words, these assessments take place per

discipline and are intended as exercises to gather input and evidence on how to maintain and further improve quality and impact of university research. Like universities in Sweden (van den Besselaar & Sandström, 2020), universities in Flanders are autonomous in the organisation of these research assessments per discipline.

In addition, universities for many years also had a legal obligation to evaluate each university professor at least every five years (recently this legal obligation has been relaxed). Although Flemish universities were in principle free to decide on how to conduct such individual evaluations, some universities set up complex quantitative systems involving, among other elements, annual performance and goal setting reviews (DORA, 2023). The reform of those systems and the fact that all Flemish universities and the Flemish Rector's Conference (VLIR) where among the first signatories of the Coalition on Advancing Research Assessment (CoARA), illustrates that each of the Flemish universities is seeking a balance between expectations for productivity and impact, and nurturing academic freedom and diversity in research. In the next section, we delve deeper into how we balance these aspects in research assessments of teams at UAntwerp.

Research assessments of teams at the University of Antwerp

In 2007, the Research Board of the University of Antwerp decided to introduce a systematic external quality assessment of its research, through a discipline-specific approach and involving site visits by external peer reviewers (Engels et al, 2013). Since then, consecutive site visits took place according to a rotating system, in which each year the research groups belonging to two disciplines have been assessed. This way, all disciplines at the University of Antwerp have been evaluated twice since 2007. The Research Board opted for a protocol which is similar to the Dutch research assessment protocol (Standard Evaluation Protocol or SEP - since 2021 renamed Strategy Evaluation Protocol). Each international peer panel presents its assessments on four criteria – quality, productivity, societal engagement & impact, and viability - according to a five-point scale: (5) excellent, (4) very good, (3) good, (2) satisfactory, and (1) unsatisfactory. The panel provides a textual motivation for each score. Next to scoring the groups, the panel also reflects and provides feedback on the research policy of the department and/or faculty to which the groups belong and makes suggestions for the further development of research policy at the level of the department, the faculty, and the university (Houben, 2023).

The stated aim of the assessments is improving the quality and the impact of the research. By assessing the performance of the groups against their mission, strategy, and future plans, the panel members provide feedback on the past performance and current situation of each of the groups and are able to provide recommendations towards the future. Each assessment is to be regarded as a guiding principle, a means of self-reflection and positioning one's team in the research system. Although the units of assessment are the research teams, the assessment is strongly related to research policy within the department, the faculty, and the university. By assessing all groups within a department or faculty, the panel gets a broader picture and can make recommendations to each aggregation level where it sees fit. After all,

difficulties in the research agendas of the groups often are related to obstacles within the research policy on a higher level. As such, each assessment report provides each level recommendations by an international expert panel about the research itself, the research context and the research policy.
Ever since 2007 the assessment dossiers prepared in view of each visit emphasize a holistic approach, diversity, transparency, and validity. In the dossiers each research team provides qualitative context in the form of their mission, strategy, achievements, and a SWOT-analysis. Quantitative measures and indicators, that are known to and validated by the researchers in the discipline and each of the research groups prior to the submission of the preparatory documents to the expert panel, support these narratives. These quantitative indicators included information on inputs (e.g. overview of academic and technical staff, as well as doctoral and postdoctoral researchers; overview of funding acquired), process (e.g. duration of doctoral trajectories), as well as output (e.g. doctorates awarded; publications; patents), and impact (e.g. citations; spin-offs launched). In terms of scholarly outputs, the approach can be considered broadly in line with the recommendations of the Leiden Manifesto (Hicks et al., 2015), e.g. the inclusion of outputs in a diversity of languages and beyond international citation databases, and decided upon in consultation with the researchers in the discipline of focus. Still, the channels of publication and, where applicable, their impact factors are provided, as are personal bibliometric profiles of the professors in the group, leading to a possibility for focus on specific kinds of outputs (e.g. publications in high impact journals) over others. Over time we have put more emphasis on societal impact and incorporated information on Open Science practices, as well as on research integrity and a diversity of research outputs. The university research affairs office also applies a co-creation approach in the assessment process, by taking into account discipline specific characteristics and needs throughout the process. Such a way of evaluating has gained more and more attention since the creation of the SCOPE Framework (INORMS, 2021). UAntwerp professors also suggested potential panel members and chairs (Rahman et al., 2016). The entire process was also carried out in a transparent way by granting the professors and researchers access to all documentation and information with regard to the research assessment, including all the details behind numeric tables and graphs that the research affairs office prepares for the international expert panel (cf. Hong Kong Principles for assessing research, Moher et al., 2020).

Observations after two cycles of research assessments

In this section we provide a brief summary of observations after two cycles of research assessments at the University of Antwerp. We specifically zoom in on three aspects:

- The correlation of the assessment scores. As research groups receive scores on quality, productivity, impact and viability, we ask ourselves to what extent these ordinal scores correlate with each other. The higher the correlations, the more difficult panels might find it to differentiate these predefined dimensions during their assessments. Very high correlations may indicate a need to limit

the number of dimensions to assess, while dimensions that are less correlated indicate areas that might be in need of additional attention.

- The evolution of assessment scores from the first to the second cycle of assessments. Houben (2023) already reported, over halfway the second cycle, a clear increase of scores. Here we analyse this evolution upon the completion of the entire second cycle, and zoom in on the evolution of the scores for each of the four criteria.
- Lastly, we consider the main recurring issues that expert panels commented on, and what they might imply for the setup of the research assessments.

Correlation of scores

We calculate Spearman's rho correlations for the assessment scores within the first assessment cycle and within the second assessment cycle. Within each cycle, one expert panel per discipline assessed all the research teams within the given field. All correlations are positive and statistically significant ($p<.001$), yet the correlations in the second cycle are lower than in the first cycle, implying more variation of the scores per group in the second cycle. Especially in the first cycle, the high correlations seem to indicate the difficulties panels may experience to assess these predefined dimensions of the performance of teams separately. In the second cycle we still observe moderate correlations, although some panels were more inclined to e.g. assess impact and/or viability differently than the quality and productivity dimensions.

Table 1. Correlations of scores for quality, productivity, impact, and visibility in the first (upper right triangle) and the second (lower left triangle) cycle of research assessments.

Table	*Quality*	*Productivity*	*Impact*	*Viability*
Quality	-	.76**	.76**	.75**
Productivity	.60**	-	.73**	.65**
Impact	.43**	.49**	-	.73**
Visibility	.45**	.53**	.50**	-

Note. ** p <.001

Evolution of scores

We performed Mann-Whitney U tests for nonparametric assessment scores to evaluate the differences between scores awarded by the international expert panels to the research groups in the first cycle (N = 136) and those awarded in the second cycle (N = 112). The results indicate statistically significant differences between the scores on all criteria: quality ($z = 4.07$, $p < .001$), productivity ($z = 4.47$, $p = .001$), impact ($z = 2.34$, $p = 0.019$), and viability ($z = 2.27$, $p = 0.027$) with higher scores on these parameters in the second cycle (Figure 1).

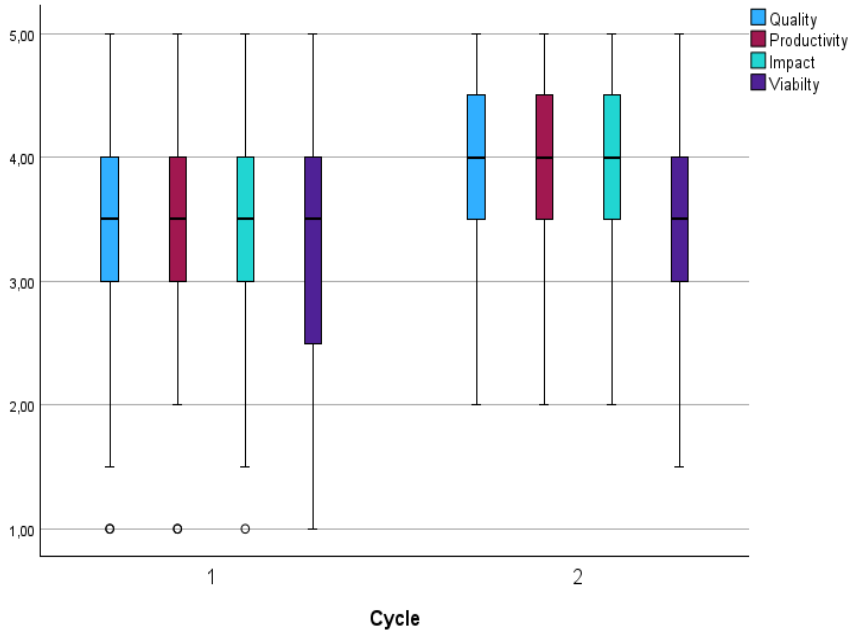

Figure 1. Scores per assessment criterion during the 1st (136 groups) and the 2nd cycle (112 groups) of research assessments.

Van Drooge et al. (2013) observed a similar inflation of scores in the Netherlands. Since the members of an expert panel in the second cycle receive the assessment report from the first cycle, they often make comparisons with previous recommendations and scores. As many groups try to live up to these recommendations, panel members tend to reward the groups with a higher score than in the previous cycle. Indeed, the intensity of research at Flemish universities, e.g. in terms of number of researchers per professor, has steadily increased at least until 2020, thus providing support for higher scores in terms of productivity. At the same time, the evidence is mixed at best when it comes to per capita productivity of research outputs, or quality and impact of research. Therefore we consider the gradual increase of scores for quality, productivity, and impact likely to represent a learning or habituation effect, whereby the teams learn to position themselves more strategically whereas the assessors reward positioning and results that are in line with their expectations and recommendations.

The increasing scores for viability, however, probably also represents another effect, that of further clustering of research groups. Indeed, even though the UAntwerp added two new faculties (in Applied Engineering and in Design Sciences), and two new departments (in Revalidation Sciences and in Applied Linguistics) in 2014, the number of research groups in the second cycle is considerably lower than in the first cycle, mainly due to mergers of groups into somewhat bigger wholes. Indeed, with the current total of 112 research groups, the average group now brings together 4 full time equivalent of professorial appointments, while this used to be 3 FTE.

Recuring themes in the self-assessments and the expert panel reports

In terms of recurring themes addressed in the expert panel reports, we distilled five main themes across all assessments. A major observation is that very few aspects of these themes seem to relate to one of the assessment dimensions specifically.

Not surprisingly, a first recurring theme concerns the attraction and retention of talent. Indeed, research cannot do without talented and well-trained researchers at all levels, from PhD candidates to professors. Hence research groups and departments often brought up this challenge, while the expert panels repeatedly stressed the importance of investing in early career researchers, their training and their career paths.

Secondly, research groups and departments frequently brought up funding for research as a theme, often with concerns regarding the high competition for grants and fellowships, and the extensive uncertainty that comes with low success rates at all levels (e.g. at the Flemish and European level). We find this also in the panel reports, in particular the recommendation to strengthen support in the pre-award phase.

Thirdly, the units frequently commented on the need for state of the art research infrastructure, including the cost of maintenance of such infrastructures. The expert panels for their part underscored the importance of research infrastructure, while also emphasizing the need for collaboration and efficiency gains, and the need to prioritize long-term investment in technical staff and infrastructures.

Fourthly, the high workload of researchers and professors is a recuring theme, attributed to a range of issues such as administrative overload and teaching load. Indeed, the expert panels recommended several times to (re)balance teaching and research, while also suggestion increasing administrative support.

We note that the expert panels tend to link these themes to the future research performance of the teams, although rarely specifically to the dimensions of quality, productivity or impact. Just like the medium to high correlations of these assessments, this seems to illustrate the impossibility for the expert panels to disentangle these different dimensions during the assessments. The same holds for other recuring recommendations, such as the suggestion to strengthen the international profile and networking of the university.

The expert panels linked only a few themes or topics explicitly to quality, productivity or impact of the research. For example, some panels recommended to aim explicitly for high impact journals, while others encouraged to focus more on societal impact. An often recurring theme was the need to facilitate and stimulate interdisciplinary research and to tackle the hurdles that researchers experience to engage in such research, which panels often linked to the innovativeness and (future) societal impact of the research.

Overall we observe that the main recurring themes in the expert panel reports, across all disciplines and across the two assessment cycles, rarely relate directly to one of the dimensions of the assessment. Rather they tend to relate to the research context, the organisation, and the (research) policies at the departmental, the faculty, and the university level.

Towards a new framework of responsible research assessments

Currently we are preparing, in close interaction with the Research Board of the university, the third cycle of research assessments. In line with the SCOPE framework (INORMS, 2021), a first focus of these discussions concerns the purpose of the research assessments. In particular, we suggest a refined notion of the purpose of the research assessments. Instead of the purpose 'to improve the quality and the impact of the research', we suggest a more specific purpose 'to contribute to a context that is conducive to high quality research with high impact on science and/or society'. This explicitly includes recognizing values of academic freedom, diversity and inclusivity of research, and contribution to the prosperity and well-being in the region and beyond.

In order to attain this ambitious goal, we intend to prioritize a thorough understanding of the context of the research as a starting point for all assessments. This will include both the broader context of research in Flanders, as well as continuous monitoring through dashboards of each research unit's performance in terms of inputs, process, outputs, and impact. We aim for more ethical and responsible use of all available indicators, by positioning them explicitly as background information to the mission and strategy of each research team and making this information permanently available to each team. This context per research team will provide the background for an analysis of strengths, weaknesses, threats, and opportunities by each cluster that is to be assessed. This qualitative SWOT analysis will then serve as the main input to a panel of experts.

In terms of aggregation level, we intend to evolve towards broader clusters, larger then one discipline or department and perhaps sometimes involving several faculties at once. Hence the research teams will become the building blocks of an assessment, rather then the units of assessment. What will be assessed, with a view of maximal alignment, is the research context, the organization, and the research policy at the level of the departments, the faculties, and the university. This 'assessment' will not be a numerical assessment, but will take the format of a set of recommendations, in order to foster collaboration, and aligned strategies at all levels, taking into account the needs of a variety of stakeholders, from early career researchers, to professors and research group leaders as well as heads of department, deans, and the rector team.

Conclusions

The University of Antwerp conducts research assessments of research teams since 2007. Over the years the approach of the assessments has evolved, e.g. gradually paying more and more attention to context and process elements. With the third cycle of assessment in preparation, we advocate a thorough rethink of the assessment approach, refocusing on the research context, the organization, and the research policy at the level of departments, faculties, and the university in order to align more with the ambitions of responsible research assessments. At the occasion of the ISSI 2025 Special Track FRAME, we will reflect on our experience of conducting research assessments since 2007 and the state of play of the preparation of a third cycle of responsible research assessments. The focus of our presentation will be the

challenge of integrating contextual quantitative indicators and qualitative SWOT analyses in the assessment of research.

Acknowledgments

We wish to acknowledge all researchers and local and international professors that have contributed to the research assessments at the University of Antwerp. Furthermore we wish to acknowledge the many colleagues in Flanders and beyond that have provided us input and reflections on how to conceptualize and organize responsible research assessments.

References

Abramo, G. (2024). The forced battle between peer-review and scientometric research assessment: Why the CoARA initiative is unsound. *Research Evaluation*, 1–8. https://doi.org/10.1093/reseval/RVAE021

CoARA. (2022). *Coalition for Advancing Research Assessment*. https://coara.eu/agreement/the-agreement-full-text/

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics*, *59*(2), 253–276. https://doi.org/10.1023/B:SCIE.0000018532.70146.02

DORA. (2023). *Case study: Ghent University*. https://sfdora.org/case-study/ghent-university/

Engels, T. C. E., Goos, P., Dexters, N., & Spruyt, E. H. J. (2013). Group size, h-index, and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. *Research Evaluation*, *22*(4), 224–236. https://doi.org/10.1093/reseval/rvt013

Engels, T. C. E., & Guns, R. (2018). The Flemish Performance-based Research Funding System: A Unique Variant of the Norwegian Model. *Journal of Data and Information Science*, *3*(4), 45–60. https://doi.org/10.2478/jdis-2018-0020

Global Research Council. (2024). *Dimensions of Responsible Research Assessment*. https://globalresearchcouncil.org/about/responsible-research-assessment-working-group/dimensions-of-rra/

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Houben, B. (2023). 15 years of research assessment exercises at the University of Antwerp: How evolution in the research landscape leads to change in the assessment practice. *STI 2023 Conference*.

IDEA Consult. (2024). *Systeemanalyse Onderzoek, Ontwikkeling en Innovatie en uitgaventoetsing van het 'Vlaams beleid in het kader van productiviteit'*. Departement Economie, Wetenschap & Innovatie.

INORMS. (2021). *The SCOPE framework. A five-stage process for evaluating research responsibly*. Emerald Publishing. https://inorms.net/scope-framework-for-research-evaluation/

Moed, H. (2017). *Applied evaluative bibliometrics*. Springer International Publishing.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, *18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

OECD Global Science Forum. (2025). *New expectations and demands from science: Rethinking research assessment and incentive structures*.

Rahman, A. I. M. J., Guns, R., Leydesdorff, L., & Engels, T. C. E. (2016). Measuring the match between evaluators and evaluees: Cognitive distances between panel members and research groups at the journal level. *Scientometrics*, *109*(3), 1639–1663. https://doi.org/10.1007/s11192-016-2132-x

Sivertsen, G. (2017). Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. *Palgrave Communications*, *3*(1), 17078. https://doi.org/10.1057/palcomms.2017.78

van den Besselaar, P., & Sandström, U. (2020). Bibliometrically disciplined peer review: On using indicators in research evaluation. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.16

van Drooge, L., de Jong, S., Faber, M., & Westerheijden, D. (2013). *Twintig jaar onderzoeksevaluatie: Feiten & cijfers* (p. 9). Rathenau Instituut. https://www.rathenau.nl/nl/werking-van-het-wetenschapssysteem/twintig-jaar-onderzoeksevaluatie

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. https://doi.org/10.13140/RG.2.1.4929.1363

Wouters, P., Glänzel, W., Gläser, J., & Rafols, I. (2013). The dilemmas of performance indicators of individual researchers-An urgent debate in bibliometrics. *ISSI Newsletter*, *9*(3), 48–53.

Zhang, L., & Sivertsen, G. (2020). The new research assessment reform in China and its implementation. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.15