Quantitative literacy for responsible research policy

Distinguished Rector Magnificus, esteemed Faculty Board, highly respected listeners,

It must have been around the beginning of June 2009, which makes it ten years ago now. I had been working at the Centre for Science and Technology Studies (known by its Dutch initials: CWTS) for less than two weeks, and had said that I would like to give a short presentation for my new colleagues. The week before I had been studying the calculation of the indicator developed by CWTS, known as the 'crown indicator'. This is a number that shows how often the publications of a researcher, research group or university are cited on average, and a number of this kind can be helpful in evaluating scientific research. The special feature of the crown indicator is that it takes account of differences between research fields in the way researchers make citations, so that a citation in mathematics, for instance, will have a different weighting than a citation in cell biology. The idea is that this makes it possible to compare scientific performances in different fields with one another.

After looking closely at the calculation of the crown indicator, I got an uneasy feeling. Is this really the right way to make comparisons between fields, I wondered. Should the mathematical formulae behind the crown indicator not be framed in a slightly different way to arrive at a fairer comparison between fields? The crown indicator had been introduced by CWTS in the 1990s[1] and it had brought the centre considerable renown. Ton van Raan, the Director of CWTS at that time, was travelling all round the world to promote the crown indicator. But was the indicator's calculation actually correct?

In my second week at CWTS I gave a short presentation, expressing my doubts about the calculation and suggesting an alternative one. I was certainly very nervous. The crown indicator was very important to CWTS. How would people react to my presentation? How much room would there be for internal discussion about this crown jewel of CWTS? After my presentation, at first there was silence. No-one responded. Then Ton van Raan spoke up and said that he was sensitive to my arguments and that CWTS should take a serious look at them. This was the start of a period for me that has now lasted ten years, in which I have learned a great deal about using quantitative information to support research policy. In the rest of this lecture, I would like to share with you some of the insights I have gained during those years.

Quantitative science studies

My chair focuses on quantitative science studies. We live in a world full of data, and the science system is no exception to this. We have data about the articles that academics publish. We also know how often reference is made to those articles, not only in other articles but also, for instance, in social media and blog posts, in policy documents, medical guidelines and patents. We also increasingly have access to the full text of articles, which gives a much richer picture than when we only have access to the metadata, such as the title and the names of the authors. There are statistics

[1] De Bruin, R.E., Kint, A., Luwel, M., & Moed, H.F. (1993). A study of research evaluation and planning: The University of Ghent. *Research Evaluation*, *3*(1), 25–41.
Moed, H.F., De Bruin, R.E., & Van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, *33*(3), 381–422.

about how much different countries spend on scientific research and about the grants given to researchers to fund their work. We also have data about the number of researchers per country or per university. There is additionally a growing quantity of data on how researchers assess each other's work, by means of 'peer review'.

Quantitative science studies concerns analysing these large amounts of data about the science system. This is often called scientometrics, or in other words: the measurement of science. Scientometrics helps us to understand how the science system operates. The ultimate goal, in my view, is to gain insights that support us in organising the system as effectively as possible. For instance, scientometrics can help us in evaluating research and allocating scarce resources. Scientometrics also offers access to information that can be useful in developing research policy, in the area of open science, for example, or the area of the scientific career system.

However, I would like to warn people not to expect too much from this data-driven scientometrics. While it is absolutely essential for many issues to have enough data available, and we are certainly very fortunate to have access to growing quantities of data about the science system, the value of these data is limited if we don't also have access to a theoretical framework for interpreting the data. All too often, such a framework is lacking, which means that data in scientometric analyses are interpreted in simplistic ways, which can sometimes even result in misleading conclusions.

This is why, in my view, quantitative science studies not only has to engage in measuring science, but must also focus equally on developing theoretical frameworks. On the one hand, scientometrics can provide these theoretical frameworks with an empirical basis, and on the other these theoretical frameworks can contribute to a more nuanced interpretation of scientometric observations. Quantitative models and computer simulations are highly promising methods for building a bridge between theory and empirical observation in this way. I will come back to this point.

Scientometrics and research evaluation

First, however, I would like to say a little more about the crown indicator. After my presentation in June 2009, I started – together with several colleagues at CWTS – to work on an in-depth comparison between the traditional crown indicator, for which CWTS had become renowned, and the alternative crown indicator that I envisaged. To make sure we did this carefully, we took our time about it. If we were actually going to switch to the alternative calculation, the arguments for this switch would have to be crystal-clear and the effects would need to have been systematically studied.

Various CWTS colleagues and I were still working on this at the beginning of 2010, when we were surprised by an article written by Loet Leydesdorff, a highly respected colleague from Amsterdam.[2] In this article he argued for precisely the same switch to the alternative calculation of the crown indicator. His arguments were of a different nature than the ones we were developing at CWTS,[3] but the conclusion was the same: the traditional crown indicator must be replaced by the alternative variant.

[2] Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, *4*(3), 423–430.
[3] Van Raan, A.F.J., Van Leeuwen, T.N., Visser, M.S., Van Eck, N.J., & Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. *Journal of Informetrics*, *4*(3), 431–435.

It is not particularly interesting here to look in detail at the various arguments that went back and forth in the discussion about the crown indicator. What is more interesting is how the discussion arose. As part of the contract research that CWTS conducts, it had calculated the crown indicator for researchers at the AMC in Amsterdam. Some of these researchers disagreed with the way the indicator had been calculated. Together with Loet Leydesdorff, they then approached CWTS about this. The discussion about the crown indicator was therefore much more than an academic debate among a small group of specialists. It was a discussion with consequences for how researchers and research groups, not only in the Netherlands but also in other countries around the world, are evaluated. Researchers' careers could perhaps depend on the outcome of this discussion, and this made it so important for some researchers at the AMC that they personally got involved.

Better indicators

The discussion about the crown indicator, and the awareness of the consequences that methodological choices can have for researchers' careers, led me, together with my CWTS colleagues, to put a great deal of energy into improving the scientometric indicators that were being used at CWTS. This resulted in something of which I am rather proud: a system of indicators that are defined in the most logical and consistent manner.[4]

These indicators for instance form the basis of the CWTS Leiden Ranking, a frequently used and freely available data source containing scientometric indicators for universities.[5] In addition to indicators that show how often a university's publications are cited, the Leiden Ranking also includes a wide range of other indicators. There are indicators that show how often universities collaborate with each other and with private sector organisations on scientific publications. We recently also added indicators that show the extent of universities' success in making the articles they publish open access, and the balance between male and female authors of scientific publications.[6]

The insights that we have developed at CWTS also show why it might be better not to use some indicators. The Hirsch index, or simply the h-index, is one example of these. The h-index, which was introduced in 2005, is widely used – as most of you will know – to evaluate researchers. At the same time, it is also widely criticised. In my view, this criticism is not always justified. Much of it is quite superficial and doesn't show what is wrong with the h-index at a more basic level. At CWTS we have been able to show that the h-index is inconsistent in the way it evaluates researchers.[7] If the h-index reveals, for example, that the researcher Jansen is initially performing better than the researcher De Vries, this can suddenly be reversed a short time later, even though the two researchers have performed exactly the same in the meantime.

Similar inconsistencies are evident in popular university rankings, which are highly visible in the media and also require the due attention of university managers. The most fundamental error made by these rankings is perhaps that they mix together indicators that depend on a university's size and

[4] Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S., & Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.

[5] Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., Van Eck, N.J., ... Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432.

[6] Waltman, L., & Van Eck, N.J. (2019, May 15). CWTS Leiden Ranking 2019 provides indicators of open access publishing and gender diversity [Blog post]. Retrieved from https://www.cwts.nl/blog?article=n-r2w2b4

[7] Waltman, L., & Van Eck, N.J. (2012). The inconsistency of the *h*-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406–415.

indicators that are independent of this. Suppose that I ask you which is the richest country in the world. A commonly used metric is the gross domestic product. If we use this metric, we see that in absolute terms the United States is the richest. However, if we take a relative perspective and look at the per capita GDP, then Luxembourg is the richest. Both of these perspectives – absolute and relative – have their own relevance. Depending on what exactly you want to know, you have to take either the absolute perspective or the relative one. What is completely illogical, however, is to try in some way to calculate the average of the two perspectives. From the technical point of view, it is possible, but it is then totally unclear how the results should be interpreted. Yet in popular university rankings it is the most normal thing in the world to mix absolute and relative statistics in this way. This makes it very unclear what these rankings actually tell us. Partly, they tell us something about the size of universities: the ones with a lot of money and a lot of researchers generally rank highly. And they also partly say something about universities' performance relative to their size. However, these two perspectives, the absolute and the relative, are mixed together, so no-one really understands exactly what these rankings are actually telling us.

Neurosurgery vs. neurology

I have now offered a few critical reservations about frequently used scientometric indicators, such as university rankings and the h-index. The discussion about how the crown indicator is calculated shows that also the indicators we often use at CWTS are not exempt from criticism. To demonstrate how difficult it is to define good scientometric indicators, I would now like to come back to this issue again.

After changing the calculation of the crown indicator on the basis of new insights, we were confronted one day by another complaint, and this time it actually came from our own university. CWTS had calculated the crown indicator for research groups at the LUMC, the Leiden University Medical Center. We were contacted by Wilco Peul, a neurosurgeon at the LUMC. We had calculated the crown indicator for Peul's research group, and the results were not particularly good. The scores for his group were relatively low. He informed us, however, that this was not a fair representation of the scientific impact of his group's work. In Peul's opinion, the low score for his group was a consequence of a specific methodological choice in the calculation of the crown indicator. He was able to explain this in detail.

The crown indicator involves directly comparing citation statistics within scientific fields. Citation statistics in different fields, such as mathematics and cell biology, are not directly compared with each other. By only making direct comparisons within scientific fields, you avoid comparing apples with oranges. This is the idea, at least. However, one very difficult point remains: what exactly is a scientific field? How do you define it?

It was precisely on this point that Peul felt the crown indicator has a methodological problem. Peul is a neurosurgeon. He pointed out, however, that neurosurgery is not regarded as an independent field in the calculation of the crown indicator. Instead, it is labelled under 'clinical neurology' and seen as part a larger field, which covers not only neurosurgical research but also neurological research. Peul told us that neurosurgery and neurology are very different fields, which largely work quite separately from each other. His view was that it is unfair to directly compare research in these two fields, especially because neurological research is usually cited much more than neurosurgical research.

Peul's criticism was forceful, and its detailed nature offered us at CWTS the opportunity to investigate step-by-step whether his criticism was justified. For this investigation, we used special visualisation techniques, which I will describe in more detail in a moment. Using these visualisation techniques, we were able to establish that Peul's criticism was correct, down to the smallest detail.[8] Neurosurgical and neurological research were lumped together, while our visualisations showed that they are indeed two separate fields. Next, by colouring our visualisations in a special way, we were also able to gain insight into citation patterns. Once again, we confirmed what Peul had said. Publications in neurology are indeed cited more than neurosurgical publications. Direct comparisons between the two research fields are therefore problematic. Moreover, the differences involved are not minor ones that perhaps could still be considered acceptable. We found that neurological research is cited on average at least twice as much as neurosurgical research. These are very serious differences, especially when you think that the survival of research groups can partly depend on the results given by the crown indicator.

Better algorithms

Peul's case contributed to all kinds of technical innovations at CWTS. Nowadays, we think more carefully about how we define scientific fields. We increasingly choose an approach where fields are defined using an algorithmic bottom-up methodology based on how researchers cite one another's work.[9] Publications that are strongly linked in terms of citation relations are clustered in the same research field, while publications with very few citation relations are classified in different research fields. The advantages of this approach are that research fields can be defined at a much higher level of detail and that the definitions of research fields co-evolve with developments in science. New emerging research fields are automatically identified by the algorithms that we use.

Over the years, we have continued to develop these algorithms further. Together with my colleagues Vincent Traag and Nees Jan van Eck, I recently published what is known as the 'Leiden algorithm'.[10] This is a greatly improved variant of the Louvain algorithm, a popular algorithm that has been used worldwide in thousands of scientific studies, in many different fields.[11] The Leiden algorithm solves some specific defects of the Louvain algorithm and is also many times faster. Using the Leiden algorithm, it can take less than half an hour to identify research fields in the whole realm of science on the basis of tens of millions of publications and hundreds of millions of citation relations. And the Leiden algorithm can also be used outside of scientometrics. For instance, we were recently told about a large-scale biological research project in which researchers at Stanford University are using the Leiden algorithm.

Large commercial organisations are also interested in the algorithms we develop at CWTS. Elsevier, for example, uses our algorithms in research analytics software that it sells to many research institutions.[12] Fast adoption in widely used products of algorithms developed at CWTS shows the

[8] Van Eck, N.J., Waltman, L., Van Raan, A.F.J., Klautz, R.J.M., & Peul, W.C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*(4), e62395.
[9] Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.
[10] Traag, V.A., Waltman, L., & Van Eck, N.J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233.
[11] Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008.
[12] https://www.elsevier.com/solutions/scival/releases/topic-prominence-in-science.

importance of our work, and will hopefully lead to constant improvement in the quality of scientometric analyses.

Contextualised scientometrics

The algorithmic work that I have now briefly described has solved some of the problems that Wilco Peul brought to our attention with his criticism of the crown indicator. A solution can also be sought in a different direction. I already mentioned that we used advanced visualisation techniques at CWTS to verify Peul's criticism. This involved using the VOSviewer software for scientometric visualisation,[13] which Nees Jan van Eck and I have developed over the last ten years. This software is freely available and is used throughout the world on a large scale. More than 500 articles in which it is used have been published. Many of you will be familiar with the visualisations, or science maps, that can be created with VOSviewer. For anyone who isn't familiar with these visualisations, I should really show you an example now. In fact, I have an example with me, on the lining of the gown that I'm wearing, but I think this might not be quite the right moment to open the gown; however, I will be happy to show you after this lecture has finished.

VOSviewer visualisations made it possible to verify Wilco Peul's criticism in detail. As I just explained, we used the insights gained from this to improve the crown indicator by switching to algorithmic definitions of research fields. However, you can also ask yourself a more fundamental question. If visualisations can expose the deficiencies of the crown indicator, could they not also offer an alternative for this indicator? In other words, should we really be trying to capture the scientific performance of researchers and research groups in one-dimensional numerical indicators? Would it not be much better to give a more pluralistic, multidimensional and contextualised picture of scientific performance?

This is a very topical discussion, for example in the context of the debate on 'recognising and valuing' that is currently ongoing in the Dutch research world.[14] At CWTS we are completely in favour of a more pluralistic approach. For instance, the influential Leiden Manifesto, of which I am one of the authors, argues for this.[15] Offering a more pluralistic approach means firstly that research must not be evaluated exclusively on the basis of scientometric parameters. For me it also means that scientometrics must be used in a different way than is now often the case. This is mainly important when scientometrics is used at the level of individual researchers or research groups. Here I argue for an approach that I call 'contextualised scientometrics'.

In contextualised scientometrics it is always possible to delve deeper into the scientometric data. You must be able to see, for example, what lies behind every number. What are the underlying data? How did we arrive at the number on the basis of those data? Visualisations like the ones we can make with the VOSviewer software offer a way to delve deeper into the data. On the one hand, these visualisations give more in-depth insights than numerical indicators, as evidenced by the case of Wilco Peul, for instance. On the other hand, they still give a simplified representation of the

[13] Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.
Van Eck, N.J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 285–320).
[14] https://www.scienceguide.nl/2019/05/elke-wetenschapper-die-te-lang-aan-de-universiteit-wordt-gehouden-is-er-een-te-veel/.
[15] Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, *520*, 429–431.

underlying data, which helps the users – for whom the underlying data are perhaps too complex and detailed – to actually see patterns in the data.

The idea of contextualised scientometrics differs from the standard view of scientometric indicators on essential points. Contextualised scientometrics does not make an a priori distinction between valid and non-valid indicators. Its aim is rather to enable users to gain a very deep level of understanding of an indicator, so that they can establish for themselves whether a particular indicator is relevant for a particular question or not. Users can also decide for themselves how an indicator can best be interpreted. Scientometricians have endless discussions about how citation statistics can or can't be interpreted. Do citation statistics reveal anything about the quality of research, do they show the scientific impact of research, or perhaps how it is used? Similar discussions go on in what we call 'altmetrics', such as statistics about how often scientific articles are mentioned on Twitter. Is this an indicator of the impact of research on society, or is that nonsense?

In contextualised scientometrics, the idea is to recognise that it is not possible, and not even desirable, to provide univocal answers to questions of this kind. Terms like excellence, quality and impact are fairly ambiguous and have a different meaning for everyone. It is therefore better if scientometricians try to avoid interpreting them too much. This should be done by the users of scientometric analyses, in a manner that suits their purposes. In contextualised scientometrics, the role of scientometricians is to give users the best possible access to scientometric data in an intelligently layered way, for example with interactive visualisations. Scientometricians do not need to engage in deep interpretations of the data. The users can do this themselves.

The concept of contextualised scientometrics, as I just outlined it, also offers a way to bridge the gap between quantitative scientometrics on the one hand and qualitative expert judgement on the other. As I see it, this gap doesn't really exist. We have all created it ourselves by taking a very narrow view of scientometrics, and we must stop doing this as soon as possible.

Open citations

Contextualised scientometrics not only involves a new way of thinking about the use of scientometrics. It also creates a need for open scientometric data sources. As long as scientometrics mainly consists of static reporting of numerical indicators, it is not really essential for the underlying data to be open. In the case of contextualised scientometrics, however, it must always be possible to delve deeper into the data. We can only do this by working with data sources that are sufficiently open.

Frequently used data sources like Web of Science and Scopus currently do not offer this level of openness. The producers of these data sources are reluctant to let them be very open, because this could perhaps harm their commercial interests. Unfortunately, this makes it difficult to fully realise the philosophy of contextualised scientometrics on the basis of Web of Science or Scopus data.

The good thing is that important developments are taking place at the moment in the direction of open data sources. A high-profile one is the Initiative for Open Citations, a lobby group that aims to promote the open availability of references in scholarly publications.[16] This initiative is supported by CWTS and in the near future it will hopefully also target other kinds of metadata, such as abstracts and affiliations, in addition to references. The Initiative for Open Citations has already achieved considerable success. In less than a year, most of the major scholarly publishers have started to

[16] https://i4oc.org/.

participate, which means that around half of all references, which is about half a billion of references, are now openly available.

It is just unfortunate that there are also a few major publishers who still do not support the idea of open citations. Important ones are the American Chemical Society, Elsevier, and IEEE. It seems that these publishers are not participating in open citations because in addition to their publishing activities they also have other financial interests that they want to protect. Elsevier, for example, has substantial financial interests with Scopus. By not participating in open citations, organisations are using their role as a publisher to protect their interests in the market for information products. Organisations seem in this respect to be concerned only with their own success, while forgetting about their wider social responsibility.

I believe that we should all try to remind publishers of their social responsibility. I took action myself earlier this year by resigning – with the support of my entire editorial board – as the Editor-in-Chief of the Elsevier publication *Journal of Informetrics*.[17] The journal was successful in many ways, also thanks to the efforts made by Elsevier, but Elsevier's refusal to make references openly available is totally at odds with the interests of the scientometric community. Together with the former editorial board of *Journal of Informetrics* we have launched a new journal, *Quantitative Science Studies*. This journal is owned by the scientometric community, via the International Society for Scientometrics and Informetrics (ISSI). This means we can decide for ourselves how we organise the journal. We have chosen, for example, a fair open access publishing model and full support of open citations.

Scientometrics and research policy

Up to now, I have spoken about the use of scientometrics in research evaluations. I would now like to broaden the discussion to research policy in general. In particular, I would like to say a little about the ambition to provide scientific evidence for research policy. To do so, I am going to present four propositions. Some of you may have a clear opinion on these propositions.

Proposition 1. Citations are usually distributed very unevenly across the articles in a journal, so it is wrong – from the statistical point of view – to use the journal impact factor to assess individual articles and their authors.

Proposition 2. The weak correlation between the assessment of articles on the basis of peer review and on the basis of citation counts proves that peer review can't be replaced by citation analysis.

Proposition 3. Scientometric analyses show that open access publication of an article leads to more citations.

Proposition 4. Analyses of publication and citation statistics show that there is a gender bias against women in science.

These four propositions are accepted by many scientometricians, and also have considerable influence on research policy. However, all four of them are based on research with serious conceptual and methodological deficiencies. I will briefly explain this for each of the propositions.

[17] https://www.cwts.nl/news?article=n-r2v294.

Proposition 1. There are good reasons to be concerned about the strong influence of the journal impact factor in research evaluations.[18] However, it is not correct to claim that this use of the journal impact factor is wrong from the statistical point of view.[19] Although it is true that the journal impact factor is not a good predictor of the number of citations of an individual article, this does not mean that it can't be a useful indicator of certain aspects of an article's quality: citations and quality are not the same thing.

Proposition 2. At the level of individual articles there is indeed a weak correlation between peer review and citation counts.[20] It should be noted, however, that the internal consistency of peer review is also low. Moreover, weak correlations at the level of individual articles do not rule out the possibility of stronger correlations at higher aggregation levels. Weak correlations at the level of individual articles are therefore not a valid reason to reject the use of citation analysis at higher aggregation levels.[21]

Proposition 3. Various studies have indeed shown that open access articles are on average cited more than articles that are not open access.[22] However, nearly all of these are observational studies that show only correlations and not causal connections. It is therefore not possible to conclude that open access leads to more citations.

Proposition 4. There are studies with a randomised design that provide disturbing evidence for the existence of biases against women in science.[23] Publication and citation statistics are also often put forward as evidence, but this is wrong. Once again, these are observational studies, which show only correlations and not causal connections. Studies of this kind often make corrections for confounding factors, but it is impossible to correct for all of them. Although these studies do indeed find differences between men and women in publication and citation statistics, they do not demonstrate that these differences are actually caused by biases against women.

Quantitative literacy

I am concerned about the deficiencies of some quantitative science studies research and the influence these deficiencies have on research policy. These deficiencies are certainly not unique to science studies. They are encountered in many other domains where policy-relevant research is conducted. With the aim of doing something about these deficiencies, I would like to argue the case for more quantitative literacy. This does not mean that we should perform more quantitative analyses or that these analyses should become more technically advanced. By more quantitative literacy, I mean that we should gain a better understanding of quantitative analyses and of the

[18] Wouters, P., Sugimoto, C.R., Larivière, V., McVeigh, M.E., Pulverer, B., De Rijcke, S., & Waltman, L. (2019). Rethinking impact factors: Better ways to judge a journal. *Nature*, *569*, 621–623.
[19] Waltman, L., & Traag, V.A. (2017). *Use of the journal impact factor for assessing individual articles need not be wrong*. arXiv:1703.02334.
[20] Wilsdon, J.R., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., ... & Tinkler, J. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. Retrieved from https://doi.org/10.13140/RG.2.1.4929.1363
[21] Traag, V.A. & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, *5*, 29.
[22] Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., ... & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375.
[23] Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), 16474–16479.

conclusions that can be drawn from such analyses. In concrete terms, I propose an approach based on three pillars: better quantitative research, realistic expectations of the research, and high-quality education. I will now discuss each of these pillars in turn.

Better quantitative research

I will start with the first pillar: better quantitative research. Unfortunately, things go wrong all too often at a fairly basic level. An example is the problematic focus on statistical significance, which leads to the results of research being presented in an unnaturally binary manner – that is to say, significant or not significant – so that the practical significance of effects that are found is totally overlooked.[24] In line with the idea of estimation statistics, practical significance must be given much more attention and statistical uncertainty must not be viewed in an unnaturally binary light.[25] Just as it is naïve to think that scientometric indicators can be used as an objective criterion to distinguish between good and less good researchers, it is equally naïve to think that statistical significance tests can objectively show which effects are important and which are not.

Another problem lies in the confusion between correlation and causality. Randomised experiments are often seen as the appropriate way to establish causality. The difficulty is, however, that it is often impossible from the practical point of view to conduct such experiments, or that experiments can only be conducted using an artificial design, which is far removed from real life. Most scientometric studies therefore have an observational design. The results of these studies are then often interpreted in causal terms without good substantiation. This is exceptionally problematic, because such interpretations need to be substantiated very carefully. There are systematic methods that can be used to arrive at causal interpretations on the basis of an observational research design,[26] but unfortunately they are almost never used in scientometrics.

A third problem concerns the way researchers formalise their theoretical ideas. In many cases, they do this with simple linear models. However, models of this kind often do not give a precise description of the researchers' theoretical ideas. We therefore need more tailor-made models to correctly formalise these ideas. In other cases, no formalisation whatsoever takes place. While this may not always be necessary, it certainly entails the risk that incomplete or inconsistent aspects of the researchers' theoretical ideas will be missed, and that logical errors will be made, resulting in wrong conclusions. To prevent this, researchers should translate their theoretical ideas into formal models more often. In some cases this will lead to models that can be studied analytically, while in others it will be necessary to use specialised algorithms. More use will also have to be made of agent-based modelling and other simulation approaches.

Realistic expectations

---

[24] Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307.
Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*(S1), 1–19.
[25] Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.
[26] Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
Hernán, M.A., & Robins, J.M. (forthcoming). *Causal inference*. Chapman & Hall/CRC.

I have now outlined how the standard of research in quantitative science studies can be improved. It would be naïve, however, to think that better research will solve all the problems. This brings me to the second pillar: realistic expectations of the research.

We have high expectations of science studies research. Some of us may even have the ambition to achieve 100% evidence-based research policy. However, this is unrealistic. In fact, it is even unrealistic to think that there can be straightforward quantitative answers to the many questions that confront us. This is precisely the problem with my two earlier propositions about the journal impact factor and the comparison between peer review and citation analysis. In these propositions, an attempt is made to answer an extremely complex question on simple, one-dimensional quantitative grounds. A proper answer to questions of this kind must take account of a whole range of relevant considerations. Some considerations can be quantified, while for others this would be difficult or impossible. All the various considerations must ultimately be weighed, and there will always be room for discussion about the degree of importance they should each be attributed. I believe that we should abandon the idea of being able to answer such complex questions with a simple 'yes' or 'no'.

Wherever possible, of course, research policy must make use of insights from science studies research. However, while we sometimes accuse people who use scientometric indicators of attaching too much importance to the 'numbers', we have to avoid making a similar error ourselves by giving simple, one-dimensional quantitative answers to complex questions. Responsible research policy is supported by a combination of quantitative and qualitative science studies research, but will also be achieved partly on the basis of more anecdotal evidence and political trade-offs. If we are realistic about what science studies research has and doesn't have to offer, we will perhaps make it easier for ourselves to resist the temptation of drawing overly strong conclusions from our research.

High-quality education

The third pillar, high-quality education, is perhaps the most important in the aim to achieve a higher level of quantitative literacy. In all kinds of domains, from the academic world to healthcare, crime-fighting and commerce, quantitative methods are used to develop policy. Universities – both research universities and universities of applied science – offer many different degree programmes in which students acquire in-depth knowledge of quantitative methods. A particularly important development, in my view, is the growth in multidisciplinary programmes, such as the data science programmes designed by various universities.

However, it seems that these programmes do not devote the requisite amount of systematic attention to the use of quantitative data-driven methods in the specific context of policy development. Given the increasing importance of quantitative methods in this area, there is a need for degree programmes with a focus on the use of quantitative methods for policy development. In such programmes, education in quantitative methods must be combined with thorough education in the social sciences, giving attention to the policy context in which quantitative methods are used. In addition, insights from the humanities can be introduced to encourage deeper reflection on what quantitative methods can and can't tell us. Programmes like this will educate students to become not so much experts in the technicalities of quantitative methods as experts in the responsible use of these methods in the context of policy development. Our society has an urgent need for this. I hope that we at Leiden University can take initiatives in this area. CWTS is very happy to contribute to such initiatives.

The crown

I started my lecture with the crown indicator of CWTS. There has been much discussion about this indicator over the years. The discussion involved all kinds of interesting technical issues, but insufficient attention may have been given to the wider context in which the indicator is used. Over time, the importance of this wider context has become much more evident. I believe it is now clear that searching for a crown indicator is not the most productive principle on which to base our work. The real crown for which we all must search is a balanced perspective on how scientometrics can contribute to constructive research evaluations and responsible research policy. I see this as the biggest challenge for the coming years. I am looking forward to working together with many of you, both within and outside of CWTS, to meet this challenge.

Words of thanks

This brings me to the end of this lecture. I would like to thank everyone who has contributed to my appointment. I would like to personally mention a number of people.

These are, first of all, my teachers at Erasmus University Rotterdam, Uzay Kaymak, Rommert Dekker and Jan van den Berg: I learned a great deal from you, and I especially value the great freedom that you gave me in my first steps as a researcher. This has been extremely valuable for me.

I was first appointed at CWTS by Ton van Raan, the Director of the centre at that time. Ton, thank you for the confidence you showed in me from the start, and for the opportunities you offered me to interfere in all kinds of things, even in your treasured crown indicator.

I was then able to develop further in the period that Paul Wouters was the Director of CWTS. Paul, thank you for all the inspiration you gave me, for your strong leadership and for your contribution to the creation of this chair.

I am now part of the leadership of CWTS myself. I would like to thank Sarah de Rijcke and Ed Noijons: I very much appreciate our pleasant collaboration.

I would also like to thank the members of the Quantitative Science Studies research group. I learn a great deal from and with you. In particular, I want to mention Rodrigo Costas, Vincent Traag and Nees Jan van Eck. Rodrigo, thank you for all the excellent work you do for the research group. Vincent, thank you for the stimulating discussions and for the innovative ideas that you introduce into the research. Nees, we are not only colleagues but also good friends. I am pleased that this still combines so well. We work together very often, and your contribution to the things we achieve together is absolutely essential.

I end with my family and family-in-law. I am very privileged to have the stable and happy personal situation that you offer me. Thank you all for this. Eline, I don't say it often enough, but you are a wonderful support for me. I appreciate it tremendously. Tim and Björn, sorry that you've had to sit there listening to your dad for so long. It's fantastic that you were so patient!

I have spoken.