Social impact @ sciences: the end of the ivory tower?

Peter A.G. van Bergeijk and Linda Johnson (eds)

3

The New Standard Evaluation Protocol 2015-2021

Jack Spaapen

Introduction

On the first day of spring of 2014 the new Standard Evaluation Protocol was presented to the Dutch minister of Education and Sciences, Jet Bussemaker. The minister was very happy with the protocol, the third in a row since 2003 when this national evaluation system for publicly funded research was first introduced. The contentment of Mrs. Bussemaker was brought about by a number of elements that are characteristic for the new SEP which is supposed to run from 2015 until 2021. The most important component of the new SEP was, in the eyes of the minister, the fact that the number of main criteria was reduced from 4 to 3, leaving out 'productivity' as a separate criterion. Bussemaker saw this as a timely answer to the growing critique, nationally and internationally, that too much focus on producing articles has perverse effects on both the quality and relevance of scientific research.

"Productivity and speed cannot be leading factors in the evaluation of science", the minister said. Less focus on productivity also means less focus on quantitative measurements, which in principle is good for the social sciences and humanities which, as a rule, favor quality above quantity: one good book may equal many articles.

The minister was also happy with the fact that in this new protocol there was room for serious attention to questions of research integrity, a consequence of some serious fraud incidents that took place in the Netherlands. But she saw this also in a broader perspective of data management, a topic that deserves to be reconsidered in the current age of the use of massive quantities of digital data.

While it is always wonderful to know that a minister is happy about what is produced by the sector, the proof of the

pudding will of course be in the eating, and the academic community will only start consuming this meal in 2015. As an appetizer, we will take a look at the architecture of the SEP and see if we can reveal the key elements of the protocol and find out the intentions behind it and how it can help the research community to do an even better job than it was already doing. This broader view is the purpose of this article.

Road to the new SEP

The SEP 2015-2021 is the third edition of the national evaluation protocol, which is renewed every six years. We will inspect the main ideas behind the SEP. Some people speak of the "Dutch approach". It is indeed rather unusual that our national evaluation system is not linked directly to the funding of research. The outcome of evaluations is used by university policy makers in a wider context in which other elements are also weighed. Finally, I will briefly go into the concept of social impact or better societal impact, a concept that, in my view, should be replaced by the concept of societal innovation.

The front page of the new standard evaluation protocol shows some ladders that reach up into the blue sky. Some may see this as a reference to "blue sky research", but that is not the gist of the SEP. The ladders are mostly white, with the exception of the tallest one, which is red.

Without going too much into the symbolic meaning of this picture, I believe it represents the idea that the Netherlands is doing a pretty good job when it comes to scientific research (the white ladders) and that we even manage to do something really excellent here and there (the red ladder). The minister likes to refer to Dutch research as being on a high plain with some very high mountain peaks on that high plain.

As mentioned above, the SEP is reviewed every six years. All the important science organizations in the Netherlands are involved in this review, the Royal Netherlands Academy of Sciences (KNAW), the research council (NWO), and all the universities, represented by the Association of Dutch universities (VSNU).

The review of the current SEP included a small international conference last year with representatives from some nearby countries (Germany, Norway and the UK). The main conclusion of that conference was that the Dutch SEP evaluation system is working very well. It has managed over the years to maintain and even improve the level of research at all Dutch universities. In particular, the flexibility of the system was highlighted as an advantage over more centralized systems such as the UK system.

Furthermore, the review used a study conducted by the Rathenau Institute on the last 20 years of evaluation in the Netherlands. One striking result of this study was that the average score research groups or institutes received in the evaluations has gone up from roughly 3.5 to 4.5 over the past decade. It remains to be seen whether this should be perceived as a sign of Dutch excellence or of Dutch cleverness in the sense that people are learning how to play the system.

Finally, the SEP review involved a number of focus groups with key people from the Dutch academic and policy communities. All the information was brought together and presented to a small committee (with some support staff) and within half a year, the new SEP was designed and accepted by the boards of all the important organizations: the academy, the research councils and the universities.

Some dilemmas

During the review process, a number of issues came to the fore. The Rathenau study, for example, concluded that over the period of these three SEPs, starting in 2003, the universities have gained full autonomy over, and responsibility for, the evaluation process. One of the consequences is that there are no direct financial consequences attached to the assessment, certainly not at the national level. This is rather different than in a lot of other countries where there is a more central organization of the national evaluation system. Another issue is that disciplinary evaluations, which used to be standard in the Netherlands, have been marginalized. Instead of a horizontal comparison at a national level, research is now mostly evaluated at a local level. Basically, the university or institution decides what is going to be evaluated, and how. If, for instance, all the faculties in humanities or social sciences decide that they want a national evaluation it is still possible, but it rarely happens, mostly for university policy reasons. The third, and maybe the most important, conclusion of the Rathenau study was the already mentioned huge inflation of the scores. The SEP used to work with five scores where 5 was the best, really top world class and 1 was the worst. The average score in the last six or seven years went from roughly 3.5 to almost 4.5. For many faculty boards, these high scores reduced the worth of the SEP because it makes it

hard to distinguish between all these highly rated groups.

Marginal changes over time

The main goals of the SEP have remained more or less the same up until now, but in this new protocol, some significant changes have been introduced. Officially, one of the main goals has always been accountability to the government, but this goal was never really exploited. The Dutch government likes to stay at a distance from the universities, as long as they have the idea that the institutions are acting responsibly when it comes to safeguarding quality and relevance. The ministry of education and sciences had an open invitation to attend all meetings of the review committee, but they never showed up. Another main goal is, of course, the broader accountability to society, which is maybe even more important than accountability to the government. This goal is now taken much more seriously than in the previous editions, but I'll come back to that later. The other main goals are the improvement of research quality, relevance and the management of research institutes. Finally, there is always a balancing act between evaluation used as a verdict - how good are you? – and evaluation used in a more strategic way – are you doing the right things to stay strong in the future? In this edition, the accent seems to shift to the more strategic questions.

Societal relevance has thus become a more important element over the years. It was not so important in the first edition, it became more important in the second, and now in the third, the idea is that there is really a level playing field in terms of the degree of attention paid to societal relevance on the one hand and scientific quality on the other. Another change that is hardly marginal, is the reduction of the four main criteria to three. Productivity has now been left out. This is partly due to the whole discussion that the Science in

To be sure, the fact that the productivity criterion is left out in the new protocol, does not mean that it is no longer important

Transition movement brought to the fore, but it is also a consequence of a broader resistance world wide – see for instance the San Francisco Declaration of 2012.[1]

The final change that I want to mention is that the review committees, which used to consist of scientific peers, now include people with other expertise, on, for example, technical applications or societal relevance. This does not necessarily mean that a site visit committee has to include external expertise, but research institutions should at least think about how to include

1 The San Francisco Declaration on Research Assessment (DORA), initiated by the American Society for Cell Biology (ASCB) together with a group of editors and publishers of scholarly journals, recognizes the need to improve the ways in which the outputs of scientific research are evaluated. The group met in December 2012 during the ASCB Annual Meeting in San Francisco and subsequently circulated a draft declaration among various stakeholders.

the broader societal interest in the evaluation process.

To be sure, the fact that the productivity criterion is left out in the new protocol, does not mean that it is no longer important. In both the first (quality) and what is now the second criterion (relevance), committees are still supposed to look at productivity. However, no longer as an end in itself, but as part of the output strategy of the group as a whole, leading to a more balanced and intelligent consideration of productivity and quality issues.

Finally, what is really new in this protocol is the issue of research integrity. There is no score there, but the review committees are asked to look at the policy of the institute regarding the subject of integrity. It has already been mentioned why that is becoming such an important issue.

SEP philosophy and architecture

What is perhaps more interesting than these changes, is the philosophy behind them. I already said something about the reasons why productivity has been left out – basically to avoid perverse effects. But the two main ideas of the SEP are that there is 1) a balance between scientific quality evaluation and societal relevance, and 2) that there is room for all fields to be evaluated according to criteria and indicators that fit best with the way the fields work. The latter idea is clearly meant to counteract the dominance in many evaluations of criteria and indicators that fit the natural and life sciences and not the social sciences and humanities. Groups are asked to write in their self-evaluation report on their performance in the two assessment aspects: scientific quality and societal relevance. They are asked to do that in three indicator categories: output, use and recognition. The SEP however does not prescribe which indicators to use. It leaves that up to the research fields. In other words, it is a bottom-up process in the sense that research fields have to find consensus about which indicators best represent the work that they are doing. There are two important ideas behind this: one is that there is not one set of indicators which is useful for all fields, the other is that the research community knows best how to represent its research production, and should thus take responsibility here. Clearly, this means that social sciences and humanities have the opportunity to develop the system in a way that suits their modus operandi.

The idea that quantitative indicators are important has not been completely discarded, but they should only be used in fields where that makes sense. It is a well known fact that a lot of the quantitative indicators have been developed in fields

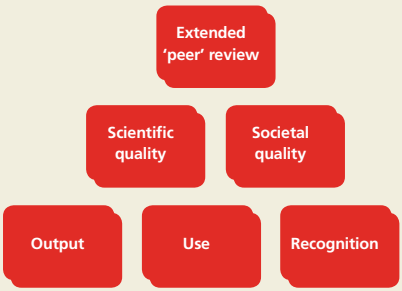
other than social sciences and humanities and that they do not work as well when applied to the social sciences and the humanities. As an alternative, the SEP offers the opportunity to write stories, narratives, that show how particular research affects society. These stories have to be underpinned with as much concrete evidence as possible. This new element, a clear reference to what is being done in the UK Research Excellence Framework (REF), is perhaps the best opportunity for researchers in the social sciences and humanities to present their work in a convincing way. More than in a lot of other scientific fields, these scientists are used to writing compelling and convincing stories. It is part and parcel of their trade.

Another point I want to raise here is the fact that the new SEP expects research groups to be aware of the policy environment. Therefore, they are asked to include in a SWOT-analysis a perspective on the surrounding policy context. This context is currently dominated by a few national and European programs, the top sectors, but also the grand societal challenges in the Horizon 2020 European Framework program. There is also the idea that universities have to look for a sharper profile, stemming from the governmental policy idea that not all universities should do the same. ''We're a small country", is the government's idea. We cannot do everything, we have to make choices.

Finally, special attention has to be paid to the review committees which conduct these SEP assessments. As a rule, these committees have a strong international signature, though the chair is often Dutch for reasons of familiarity with Dutch science policy. But now, attention should also be paid to the broader impact of research. In other words, room should be made for representatives of relevant stakeholders in the evaluation procedure.

This all leads to the following architecture for the SEP 2015-2021

Figure 1 SEP architecture



New responsibility for research fields

The idea is that in each of the three categories in Figure 1 (output, use and recognition), indicators are to be developed bottom-up by the research fields themselves. It is a very interesting and innovative idea, but how does it work? An example can be found in three reports that the Dutch academy has produced in recent years and which formed an important input into the new SEP. Interestingly, these reports were created in three different fields: humanities, social sciences and engineering and design but there turned out to be a lot of similarities across these fields.[2] The three committee chairs were able to present to the committee that designed the new SEP, a common view on how to deal with the issue of indicators (see Figure 2). Without going too much into the similarities between these fields, it is clear that these fields communicate and produce research in rather different ways than the natural sciences and the medical fields. The focus on the societal context is, for example, much stronger, and the

It is a well known fact that a lot of the quantitative indicators have been developed in fields other than social sciences and humanities and that they do not work as well when applied to the social sciences and the humanities

production of other output than articles in high impact journals is more important (for example books, experimental models, exhibition catalogues).

Differences between the schemes developed by the Academy committees responsible for the social sciences and the humanities reports are minor, differences between these two and the scheme from the engineering and design fields are slightly larger. But the basic approach in all three fields rests on the same principles: a balance between scientific quality and societal relevance and freedom for the fields to devise the indicator categories for each of the two criteria. It is important, of course, to have an evaluation committee that is sensitive to the production and communication practices in the field. Such a committee has to be able to find the right balance between scientific quality and societal relevance. Therefore, it is wise to consider involving stakeholders from the context of the research being evaluated.

2 https://www.knaw.nl/nl/actueel/publicaties/towards-a-framework-for-the-quality-assessment-of-social-science-research; https://www.knaw.nl/nl/actueel/publicaties/quality-indicators-for-research-in-the-humanities; https://www.knaw.nl/nl/actueel/publicaties/quality-assessment-in-the-design-and-engineering-disciplines.

Figure 2 Indicator scheme (examples of indicator categories)

	Scientific quality	Relevance to society
Demonstrable output	Scientific articles (refereed vs. non-refereed) Scientific books Other research outputs (instruments, infrastructure, datasets, software tools, designs) Dissertations	(policy) reports Articles in professional journals Other output (instruments, infrastructure, datasets, software tools, designs) Outreach activities, public lectures, exhibitions
Demonstrable use	Citations Use of datasets, software tools, etc. by peers Use of research facilities by peers Reviews in scholarly journals	Patents/licenses Use of research facilities by societal partners Projects with societal partners Contract research
Demonstrable recognition	Scientific prizes Personal subsidies Invited lectures Membership of scientific committees, editorial boards, etc.	Public prizes Valorisation funding Positions paid for by public parties Memberships of public advisory bodies

Research fields are thus required to come up with suitable indicators in the three categories in the above scheme. In the scheme, which appears in the SEP 2015-2021, examples are given in each of the three categories, for each of the two main assessment aspects. To be sure, these are indeed just examples. Fields remain free to make different choices, as long as there is consensus in the field, preferably through some kind of authoritative body or procedure. The idea is to trust researchers, if possible, together with relevant stakeholders, to come up with indicators that really represent their work and for which they can collect robust data, which are not necessarily quantitative data.

The question is whether this bottom up idea will work in practice. Of course, it is more easily said than done, because not only do you have to have some kind of authoritative body in a discipline or field, you also need, after you have reached consensus, the means to develop such indicators. And certainly for the social sciences and humanities, there is no organization in the Netherlands that has a lot of experience with this. It is true though that the Centre for Science and Technology Studies (CWTS) in Leiden is currently changing its course from an institute mainly focusing on traditional bibliometrics and thus natural and health sciences, to an institute with a broader focus that includes social sciences and humanities. Also, the deans of humanities have started their own project to develop new indicators.

Clearly, there is quite a long way to go. If you want the usual indicators, like the indicators that are dependent on Web of Science publications, there is not very much that you have to do. There is a lot of agreement on how to deal with that. There is also a lot of critique there too, but it has an established history and you can deal with it. If, however, you have to come up with new indicators there are quite a few steps to be taken. Take, for example, book chapters. There will be discussions about what counts as a book chapter. There will also be discussions about how to deal with the publishers, because some are more highly valued than others. There are different ways to organize peer reviews, some more and some less independent of editorial boards.

Furthermore, if you want to say something about quality, you have to have an idea of the ranking of the different media (publishers) where these articles, books or chapters appear. However, all this is still relatively easy compared to the development of indicators for societal relevance. I shall come back to that further on.

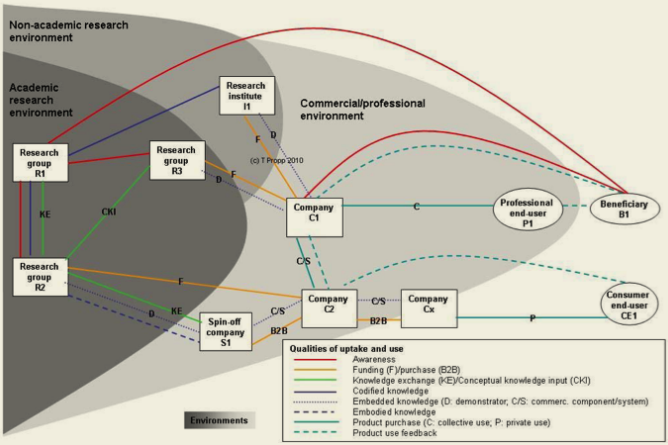
SEP in the context of policy and society

Clearly, the world outside research is changing. I think awareness of this process started a long time ago in the Netherlands. In 2010, a committee chaired by Professor Veerman produced an influential report[3], which called for more institutional differentiation, an idea that was taken up by the government in a policy paper about two years ago. Universities were then asked to write papers that showed how they were going to diversify in the near future. Another important development is the top sector policy. Basically, the government selected nine economic sectors that were seen as vital for the future of the Netherlands. Think of agriculture, chemistry, high tech, health, mostly areas where natural and health sciences are active. For the humanities there was the top sector creative industry, for the social sciences, some of them at least, the sector logistics. The government expects that there will be a growing collaboration between the research community, industry, public organizations, government organizations, and societal organizations, depending on what is at stake. Partners should show commitment by putting in financial or human resources.

Inside research, the world is changing too. The Science in Transition movement is probably the most prominent actor, alerting the academic community – researchers and governments – to the dangers of the current system. Also outside of the Netherlands there is a growing movement against the more traditional approach to quality, which is very much connected to publishing in high-ranking journals. Of course, the subject of research integrity, at least in the Netherlands and in some other countries, has also been rather prominent in the past few years. Then there is the growing attention for what I call 'MIT' research, not the famous Boston institute, but a term referring to multi-, inter-, and trans-disciplinary research. Whatever you may think of the top sector policy, it is an interesting idea to have these different orientations try to work together to address grand challenges in society. Then, as mentioned before, valorisation is an issue that is gaining prominence. It appears to be a very Dutch concept. In other countries people do not know what it means. However, if you start explaining the Dutch connotation, it becomes clear rather quickly that similar movements are developing elsewhere too. A last issue to mention here is something coming up now in European circles. It has a new acronym, RRI. It stands for responsible research and innovation. There is an official program in what used to be called science and society, but now it is

3 http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2010/04/13/advies-van-de-commissie-toekomstbestendig-hoger-onderwi.html.

Figure 3 Research and Innovation network in nano-research



© Tilo Propp

called SwafS, which stands for science with and for society. The program embraces the following 6 issues; ethics, public engagement, gender equality, science education, open access and governance. The EU expect research proposals to address these topics in applications for the Horizon 2020 program and beyond.

Societal relevance of research

I want to end with developments regarding indicators for societal relevance of research. There are quite a few projects that are working on societal relevance indicators in this new context. A few were mentioned above. Here I want to zoom in on a European research project I led a couple of years ago, www.siampi.eu.

We looked at the interactions between stakeholders in a number of different fields from the social sciences and humanities, but also from the natural sciences and engineering. Interactions were divided into three broad categories: between people, through media, and material and financial interactions. The diagram above comes from nano research and represents the complex pattern of exchanges between various stakeholders. There is also the time perspective, so starting from the original idea (left side of the diagram) to a product that consumers can use might take 10 or 15 years in some fields (right side of the diagram). During this time there are all kinds of interactions in a network of frequently changing stakeholders. To capture societal impact is

like trying to shoot at a moving target. Evidently, in this perspective, societal impact is a very inadequate concept, because it represents the idea that there is somewhere a sender and somewhere a receiver. That is a linear model and that is not often the case. On the contrary, research and innovation frequently takes place in a very interactive process. The participants and goals may shift over time. Perhaps, the long term goals remain somewhat the same, for example, clean energy or in the Intergovernmental Panel on Climate Change (IPCC), the climate goals. But to get to these long term goals, a long and winding road has to be taken. It requires input from different kinds of knowledge and expertise, combinations of natural science research and social science research. In short, it is a rather unpredictable process.

What the diagram shows is that narratives might be a better way to describe what is going on in the interactions between academic researchers and other stakeholders in the environment, often still a black box. Through these narratives, a clearer picture might be presented of what is going on in innovation trajectories. It might also be a way to think in new ways about indicators. In the SIAMPI case studies, we discussed all these things with people from the various areas that we did research in and with stakeholders in these areas. In the end they came up with these kinds of indicators. Again, you will have to do a lot of work to get really concrete indicators. Maybe that will be my final message, "We still have a lot of work to do".

Research and innovation frequently takes place in a very interactive process. The participants and goals may shift over time.