Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Effects of Metrics in Research Evaluation on Knowledge Production in Astronomy | A Case Study on Evaluation Gap and Constitutive Effects

Julia HEURITSCH

German Centre of Higher Education Research and Science Studies (DZHW), Humboldt Universität zu Berlin, Research Group "Reflexive Metrics", Department of Social Sciences, Berlin, Germany

Abstract

For this case study nine interviews were conducted with astronomers from Leiden University. The interviews were complemented by a document analysis on relevant institutional (self-) evaluation documents, annual reports, and CVs of the interviewees. The aim was to perform a qualitative study about what astronomers define as research quality and how that related to their perception on what is measured by metrics used in research evaluation. The research shows that astronomers are realists who define scientific quality on the basis of "truth" and are driven by curiosity. These two factors make up their intrinsic values and motivation to perform Astronomy. Publication pressure, arising from the requirements of "the system", creates an extrinsic motivation to perform. This is perceived as resulting in low readability and replicability, risk aversion and a focus on quantity rather than quality. Hence, indicators do not merely describe quality, but also co-constitute what counts as good research. While such constitutive effects of indicator use on research behaviour and content are observed, there is no indication that the astronomer's intrinsic values are co-constituted. This gives rise to a discrepancy between what is being measured by indicators and what astronomers define as scientific quality; the so-called 'evaluation gap'. Consequently, astronomers try to manage a balancing act between their intrinsic values and the requirements of the system. Findings on constitutive effects and the evaluation gap in Astronomy lay out the conceptual groundwork for further empirical research and for policy advice on alternative evaluation practices and innovative indicators with the aim of bridging the 'evaluation gap'.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

1 Introduction and Literature Review

The quality of scientific production is currently measured and evaluated by a set of quantitative metrics, so-called indicators. This practice has become ever more controversial as insights from the sociology of quantification and the sociology of evaluation show that numbers which quantify quality, do not only describe, but also prescribe. Indicators are performative insofar that they do not merely measure whether science is performed well, but that they also affect what is valued as good research. Reflexive metrics is a relatively new field in science and technology studies, which combines the two strands, the sociology of quantification and the sociology of evaluation, in order to study what effects indicator use has on research and researchers themselves (e.g. Stephan, 2012; Fochler & De Rijcke, 2017). Given that metrics are non-detachable from a social context, reflexive metrics will provide theories about the meaning, reliability and effects of indicator use in evaluation procedures. It is important to question the use of quantitative measurements in evaluation processes as an established practice in order to inform policy makers what effects their policies have on science and what they need to consider to encourage quality research and motivated researchers. This paper will first give a brief introduction to the topic of reflexive metrics, which roots in the sociology of quantification and (e-)valuation in *Section 1*. This includes two concepts developed to explain what effects indicator use has on knowledge production processes, the *evaluation gap* and *constitutive effects*, and how they could be reconciled. The introductive section ends with explaining why this study chose Astronomy and Leiden Observatory as the field and institute under investigation. *Section 2* outlines the methods. *Section 3* contains the results where we first depict the astronomers' definition of quality. We then describe the evaluation gap in Astronomy and what constitutive effect we could observe and finally, how those concepts can be reconciled. This paper will end with its final *Section 4*, the conclusions.

1.1 Sociology of Quantification and (E)-Valuation: Insights into the different characteristics and meanings of numbers

The question how to measure and ensure high-quality of knowledge production has become controversial and challenging. "*Accountability*" and "*transparency*" are becoming ever more closely associated with producing and monitoring metrics (Espeland &

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Vannebo, 2008). This is because quantification is one means to constitute social entities as things that last and are comparable. Categorising and numbering reduce the complexity of phenomena, which makes them easier to grasp and talk about. As such, the goal of quantification is to enable objectification and to master uncertainty. Through objectification, both a political space and a measuring space, are co-constituted in which things can be compared (Desrosieres, 1998). “It permits scrutiny of complex or disparate phenomena in ways that enable judgment” (Espeland & Stevens, 2008). Hence, quantification offers a shared language and replaces trust in people with “trust in numbers” (Porter, 1995).

Indicators serve the purpose of accountability, which is why they are relevant in science evaluation systems. Indicators *commensurate*, which is the act of using numbers to rate and rank, “creating a specific type of relationship among objects” (Espeland & Stevens, 2008). They are argued to measure and compare the output and performance of researchers and research fields. According to Godin (2006), this is also the reason why psychologists used bibliometrics as forerunners in the early 1900s. Their aim was to contribute to the advancement of psychology by demonstrating its usefulness and productivity quantified in indicators. Advancement of a research field is possible due to more positive attention from funders and policy makers, achieved by trust in numbers. That is how indicators may acquire the power to influence how funding is allocated. They are political means, solidifying categories “by means of which society seeks to manage itself and thereby represents itself and its values” (cited from Dahler-Larsen, 2014; referencing Vestman and Conner, 2006 & Rosanvallon, 2009).

Commensuration is one of the “most consequential uses of numbers” (Espeland & Stevens, 2008). Commensuration turns describing numbers into prescriptive ones. Commensuration attributes meaning to numbers. “Measures that initially may have been designed to describe behaviour can easily be used to judge and control it” and hence, “numbers can also exert discipline on those they depict” and “disciplinary practices define what is appropriate, normal, and to what we should aspire” (Espeland & Stevens, 2008). Foucault (1977 & 2003) links statistical practices to “*governmentality*”, a term to describe how the government uses numbers to influence citizens so that they fulfil those government’s policies. He describes discipline as “a mode of modern power that is continuous, diffuse and embedded in everyday routines” (Espeland & Stevens, 2008).

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

1.2 The Evaluation Gap

The fact that indicators commensurate, where “all difference is transformed into quantity” (Espeland & Stevens, 2008), leads to the argument that their use to assess scientific quality gives rise to an “*evaluation gap*” (see Fig. 18). This is a term coined by Wouters (2017) to acknowledge a discrepancy of what is being measured by indicators and the quality of the scientific content, as perceived by the researchers of the field. The researcher holds a different notion of quality than the indicator serves. The evaluation gap can lead to a number of questionable practices, such as goal displacement, gaming or information overload (Laudel & Gläser, 2014; Rushforth & De Rijcke, 2015). Because “measures can also alter relations of power by affecting how resources, status, knowledge and opportunities are distributed” (Espeland & Stevens, 2008), researchers may need to comply with the concept of quality implicit in the measurement (goal displacement). To reach the target set by the indicator the researcher may then take short cuts (gaming), which possibly undermine research quality, but fulfil quantitative requirements to publish (causing information overload). Negative effects of the evaluation gap on research practices are called “*unintended consequences*” of indicator use. This term is found frequently in literature on effects of performance measurement (for a list see Dahler-Larsen, 2014) and it draws back to the notion of “unanticipated consequences of purposive social action” (Merton, 1936).

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

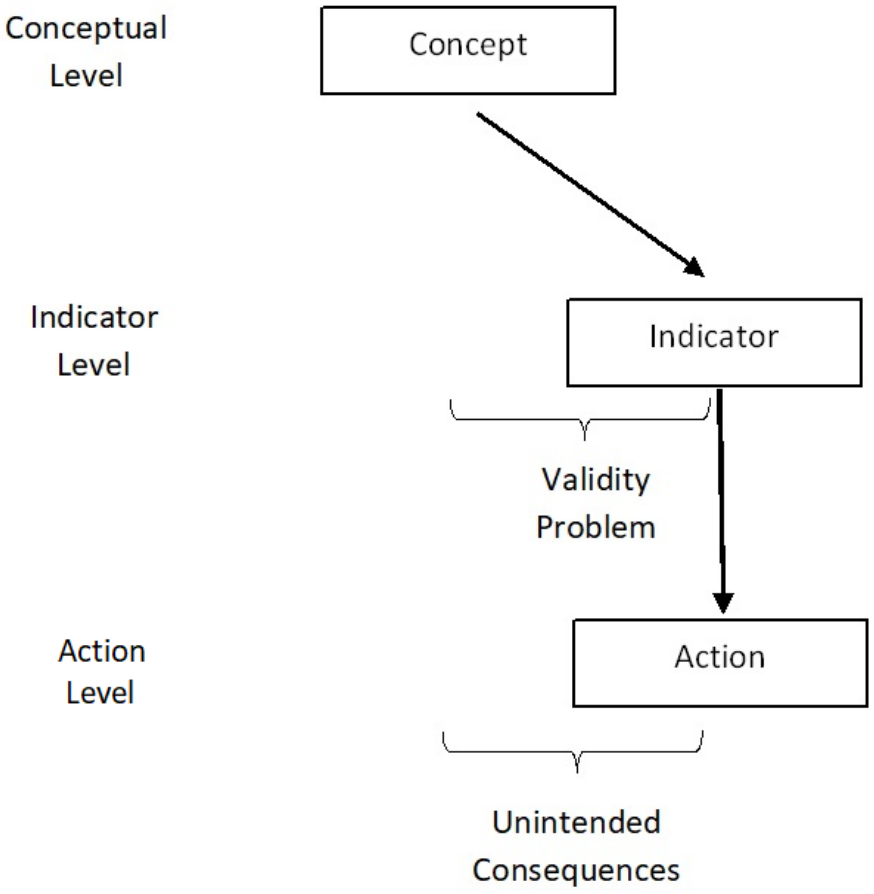

Fig. 18: *The evaluation gap as depicted by Dahler-Larsen (2014). In his paper it is called "Trivial Measure Fixation", where "the indicator is an imperfect measurement of the concept [in this paper: research quality] that is intended to measure. Despite the "validity problem" the indicator guides the action of the researchers. Due to the validity gap unintended consequences occur on the action level; the requirements of the indicator are trying to be satisfied instead of the scientists' concept of research quality.*

1.3 Constitutive Effects of Indicators

Numbers have authority and objectify, however "doing things with numbers" (Espeland & Stevens, 2008) entails a performative element. Austin's "speech act theory" ([1962] 1975) describes a specific type of utterances, so-called speech acts, that relate saying something with performing an action as such. Hence, speech acts do not simply evaluate the truth content of a statement, but they constitute an act. Analogous to that, to quantify something is always to do something, when meaning is attributed to the resulting numbers, rather than simply stating their truth content. This opens up the dichotomy between the prescribing and describing function of numbers (Desrosières, 1998). The term

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

performativity of numbers was established in the economic sociology and in the sociology of finance (Callon, 1998; MacKenzie, 2006) to convey that statistics may not only describe social realities, but also co-constitute them. The process of turning “qualities into quantities creates new things and new relations among things” (Espeland & Stevens, 2008). “Measurement intervenes in the social worlds it depicts”, as measures are *reactive*; “they cause people to think and act differently” (Espeland & Stevens, 2008).

Dahler-Larsen (2014) suggests to depart from speaking about “unintended consequences” of indicator use and using the term “*constitutive effects*” instead (see Fig. 19). On the one hand, this conceptual move avoids the “dependency on a valid identification of intentions behind the indicators” and on the other hand it acknowledges the performative character numbers can entail. Effects of indicator use are constitutive insofar that indicators are not merely representative measures of scientific quality, but they rather shape what is considered to have value in knowledge production and therefore may exert an effect on research behaviour and content. They constitute a “reality that is put on stage so that it can be acted upon” (Desrosiéres, 1998; in Dahler-Larsen, 2014) and indicators become “the way through which the world is defined” (Dahler-Larsen, 2014). For research this means that “indicators and rewards introduced by policies shape the process of the practices they seek to describe” (Dahler-Larsen, 2014). Hence, one may assume that indicator use affects research agendas, knowledge production processes and research behaviour.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

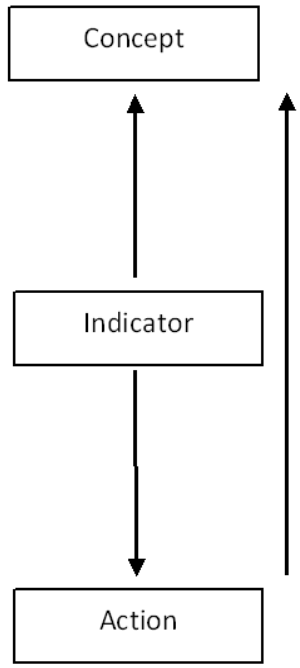

Fig. 19: *Constitutive effects as depicted by Dahler-Larsen (2014). In his paper he calls it "Advanced Measure Fixation", where indicators stand in "a constitutive relation to the reality they seek to describe". As compared to Fig. 18 there is no gap as the indicator-guided action re-shapes the concept of quality the researcher holds.*

1.4 Reconciling the Evaluation Gap and Constitutive Effects

Intentions always play a role in the processes of defining indicators and deciding on which ones to use, no matter whether they are explicit or implicit and whether they are applied in the intended way or not. By avoiding problematic assumptions about intentions, Dahler-Larsen describes constitutive effects as something that occurs to passive actors. While the evaluation gap can be criticised for not accounting for the reactivity of indicators and their potential effects on the researchers' concept of quality, the concept of constitutive effects does not leave any room for divergent notions of quality. Analysing the usefulness of both concepts raises the question whether they are necessarily alternatives or whether those two concepts can be reconciled. Dahler-Larsen (2014) recommends that "constructivists may be immediately comfortable with the idea [of constitutive effects], while rationalists and functionalists may still find value in the idea of unintended consequences." We question whether the decision which concept to use really "hinges on paradigmatic foundations" and rather propose that, when studying the reflexivity of metrics we might find

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

value in reconciling both concepts. That reconciliation may take into account that indicator use can co-constitute concepts and values (about quality), but at the same may also set targets which diverge from intrinsic values (about quality). Testing and conceptualising this hypothesis will be part of this paper.

1.5 The case study: Why Astronomy and why Leiden Observatory?

Despite Reflexive Metrics being a relatively new field, there have been quite a few studies discussing the effects of indicator use on scientists, especially in bio-medicine and the life sciences (e.g. Hammarfelt & De Rijcke, 2014; De Rijcke et al., 2015; Rushforth & De Rijcke, 2015; Fochler et al., 2016; Kaltenbrunner & De Rijcke 2016). While life science is a more applied research field, for this study we chose a field that conducts mainly basic research in order to study effects of indicators when applications don't play a significant role.

Astronomy (synonymous with Astrophysics) is one of the oldest sciences and according to Heidler (2011) there are 15,000–20,000 active professional astronomers worldwide. The author characterises Astrophysics as "a paradigmatic, established, basic, hard knowledge field with relatively clear disciplinary boundaries", often following Karl Popper's ideal of trying to develop theories that can be falsified. Paradigmatic and hard science fields typically are strongly reputation oriented, but nevertheless reflect on bibliometric measures of reputation, like the h-index (Heidler 2011). The reputation system in Astronomy is still based on individual achievements, while an increasing collaboration culture puts pressure on that system. At early career stages, decisions about who to fund and to hire, are "essentially predictions about an individual's future achievements" (Kurtz & Henneken, 2017).

Astronomy is an interesting field to study from a meta-perspective. Astronomy asks highly fundamental questions which inspire both scientists and the public at large. It is dedicated to basic research, involves large collaborations on expensive instruments such as telescopes, and the use of (open) archives and huge datasets. The access to telescopes is generally not exclusive, although the builder (and collaborators) of the telescope usually get a share of guaranteed telescope time. Instead, the access is regulated by a peer-review system, which evaluates the prospective quality, originality and success of the project (Heidler 2011).

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

According to Heidler (2011), the social structure of a field is influenced by both, its knowledge content and its "historically grown organizational and cultural preconditions". While Heidler (2011) performed a study on "cognitive and social structure of the elite collaboration network of astrophysics" and Kurtz & Henneken (2017) performed a "40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics", showing the capabilities and limitations of each measure, no study has been performed on what effects the use of metrics in research evaluation has on the knowledge production process in Astronomy.

We chose to study evaluation processes at Leiden Observatory (Sterrewacht; Leiden Astronomy institute), since it is viewed as one of the largest and top astronomical research institutes in the world. In 1998 the national Astronomy proposal "Astrophysics: unravelling the history of the universe" was rated first by the Netherlands Organization for Research (NWO). This proposal was submitted under the umbrella of the Nerderlands Onderzoekschool voor Astrononomie (NOVA). It is the alliance of the four university Astronomy institutes in the Netherlands – the Universities of Amsterdam, Groningen, Leiden and the Radboud University Nijmegen – and was rated as top research school in 1998. As a result of the proposal NOVA was guaranteed baseline funding from the Dutch Ministry for Education for 1999 to 2005. Since then this "NOVA grant" has been renewed every 5 years. The grant has been the basis for support of "normal" research activities and the participation in numerous programmes for the construction of astronomical instrumentation. This enabled Leiden to build on its long tradition of radio-interferometry, by getting heavily involved with instrumentation for European Southern Observatory's (ESO) facilities, securing priority access for conducting observations. Additionally, Leiden hosts the world famous Sackler Laboratory that bridges Astronomy, physics, chemistry and biology. Leiden Observatory is an international environment; many students, postdocs and staff come from abroad. The institute has close collaboration ties with other Astronomy institutes in Europe and the U.S. and hosts visitors from across the globe.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

2. Methods

The use of metrics to evaluate science opens up a whole series of topics that have found attention in the sociology of quantification. From the way how indicators can provide accountability and enable governmentality to different notions of reactivity, such as unintended consequences or constitutive effects. Espeland & Stevens (2008) point out that “the capacity of measures to discipline” is another “distinctive form of reactivity”. Performativity of numbers with attributed meaning is the reason why studying the effects of indicator use is relevant in the social sciences. This paper studies the extent of the indicators’ influences on knowledge production processes and research behaviour in Astronomy, including the following research questions: What (perceived) effects do indicators have on the quality of knowledge content? Can the influences be described as unintended consequences or rather constitutive effects? Or must those two concepts be reconciled? What is the relationship between the evaluation gap and constitutive effects?

In this study we classify “research evaluation” as any kind of evaluative process or situation that an astronomer is faced with, which is important for them to continue their research or their career. The career system, funding system and publication system are regarded as different aspects of the evaluation system.

The research questions were tackled by conducting interviews and a document analysis at the Centre for Science and Technology Studies (CWTS) under the supervision of Sarah de Rijcke. Since it regards itself as an elite institute and is listed among the best astronomy institutes in the world, the Sterrewacht makes a good case study of what effects research evaluation has on the knowledge production process and research behaviour. The author graduated from Leiden Observatory in 2015, so she had easy access to the institute. Ten researchers were invited to be interviewed for this study via email. This sample was selected such that it includes scientists in different career stages and from a variety of nations. The Master programme at the Sterrewacht is very research intensive, requiring the students to write two Master theses in total, which is the reason why they are also interesting subjects for this study. From the ten researchers, nine replied positively, which led to semi-structured interviews with four faculty members, two postdocs, one PhD candidate and two Master students.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

In order to investigate a potential evaluation gap, questions were developed such that an astronomer's definition of quality versus what is measured by indicators can be studied. Topics included career steps, project funding, exposure to assessments, research evaluation, the publication and funding system, different stages of the knowledge production process – from planning, via doing the research to publishing – and the meaning of quality. Each topic was introduced by one overarching question, followed by several potential follow-up questions.

Subsequently, all names were pseudonymized. All interviews, 80-100 minutes in length, were fully transcribed into electronic form, coded and summarised. These codes represent themes which emerged by combining sensitivity towards existing literature on constitutive effects of indicator use with insights from our data. As for the investigation of the astronomers' notion of quality bottom-up coding was applied. To study what constitutive effects indicator use has on Astronomy, top-down coding was done on the basis of five domains of constitutive effects which Dahler-Larsen (2014) carved out. The interview questions and codes can be found in *Table S-1 & Table S-2* of the supplementary material.

The interview data were complemented with a document analysis of materials collected online or made available via the informants, including CVs of the interviewed researchers, annual reports and (self-) evaluation reports of the Dutch Astronomy institutes and their umbrella organisation NOVA. The annual public reports were authored by the respective director of the institute and the collection used in this research comprises those written for the years 1998 to 2015 (hereafter; Annual report^{1998}- Annual report^{2015}). Institute evaluation protocols for the evaluation period 2010-2015 (hereafter, Evaluation protocol$^{2010\text{-}2015}$) were authored by an external committee and self-assessment protocols (same period; hereafter, NOVA self-assessment$^{2010\text{-}2015}$ & LU self-assessment$^{2010\text{-}2015}$) were written by NOVA and the institute as a preparation for the evaluation. Those (self-)evaluation reports are particularly interesting as they compare the Sterrewacht with their national and international counterparts and explicate by which standards successful research is measured in Astronomy. Comparing the official documents with the interviews gives insights into what is valued by evaluation practices as compared to what astronomers value in doing their everyday research. Hence, evaluation documents can help identifying an evaluation gap and constitutive effects.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

This article summarises the results while the complete report of this project can be found on ArXiv[1]). Direct quotes of the interviewees will be given between double quotation marks.

3 Results

3.1 What is scientific quality for an astronomer?

In order to understand the extent of a potential evaluation gap and how indicators shape knowledge production in Astronomy, we must investigate the intrinsic values and general motivation of astronomers and compare this with what is required by the evaluation system. Only if we know what research quality means for an astronomer, we can investigate whether indicators have constitutive effects on quality.

Our document analysis gives insights into various strategies on how Leiden Observatory maintains its "success" and how that success is evaluated and measured, both in qualitative and in quantitative terms. What we are missing from the reports is an answer to the question who defines quality and if the described measures can satisfy that definition. NOVA claims "*the first part of its strategy [to ensure a front-line role in Astronomy] is to foster an intellectually rich and vibrant scientific atmosphere which allows astronomers to pursue their ideas and push scientific boundaries, and in which young scientists can develop and grow*." This sounds great in theory, but we question, whether individual researchers feel that "success" as defined in the evaluation protocols actually allows such a "vibrant scientific atmosphere" and out-of-the-box thinking in practice.

The study found that astronomers generally conduct science for the sake of "curiosity" and "pushing knowledge forward" (e.g. PhD Candidate, Faculty Member 4), that is searching for the truth and discovering structures of nature. Astronomers are realists, who assume a reality independent from the observer, arising from (physical) causal laws. As Astronomy is the study of the universe and its building blocks, it seeks to answer the most 'fundamental' questions to set a basis for the 'truth of everything'.

1. *https://arxiv.org/abs/1801.08033*

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> "But that moment of – you know – mystery, that is a scientific experience in the sense that there is only one thing that you accept in that moment, that's the *truth*, you want to know the *real answer*. And no excuses, only the real answer matters. And that is what *drives science*; we only want to know the real answer." (Faculty Member 1)

The notion of truth and the quest to push knowledge forward both result from the astronomers' curiosity to understand the universe, which all interviewees uttered as their motivation to become an astronomer. Astronomer's intrinsic motivation is to "know and to understand better" (Postdoc 1).

> "I mean [my driver is] the journey and not the arrival, basically. […] It's just simply that it feels good. And in German they have a word for that, they call it the 'Aha-Erlebnis', the 'Oh, is that so'-feeling." (Faculty Member 3)

From the astronomers' notion of truth and their motivation to discover follows that high quality in research means that there is a correspondence between reality and the scientific theory (also compare with citation of Faculty Member 1 above). For a realist, truth and scientific quality are 'objective' and it implies scientific integrity.

> "I think in terms of what constitutes *good science* and what is *academic integrity*, all those things don't change – they are pretty close to *absolute values* I would say." (Faculty Member 2)

However, what does an astronomer define as a discovery? The research has to "be something new" (Postdoc 1) to push knowledge forward. This ranges from "trying to solve a problem, no matter what the problem is" (Postdoc 1) to "asking an important question" (Faculty Member 1) and having the means to solve the problem. Solving those problems doesn't only serve the astronomer's intrinsic motivation, but also, in their view, has a high relevance for society and other academic fields.

> "The inspiration that Astronomy brings and the fundamental questions it raises about the nature of everything and the place of humanity in the universe, makes it natural for us to engage with fellow intellectuals in seeking connections between arts, humanities, and science." (NOVA self-assessment[2010-2015])

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> "Science that drives the [knowledge] forward, is science that serves society."
> (Faculty Member 3)

The interviewees display consensus about the importance of Astronomy with respect to this mission. However, when asked for a more objective definition of what an "important question" is, the astronomer admits controversy:

> "That is […] difficult to answer, because if you have 5 referees, they will all have different preferences for what is important and what is not important."
> (Faculty Member 1)

However, astronomers do agree that there is a difference between "making progress on an important issue" and "valorisation":

> "Well, academic quality I think has always been relatively clear. It has to be verifiable and clear, unbiased etc. I think that is academic quality. But there is these days … a tendency to look at the value of science in terms of economic output, it's called 'valorisation'. And I am totally uninterested in that I have to say. It is nice if you can […] use some things… It is always nice if you find applications that are useful and that can actually make you profit even. Why not? But that's not why we do it. And the importance of that is overstated these days. And I don't think that is actually productive." (Faculty Member 1)

Here we can see again the high value of "truth" for an astronomer. Truth matters for its own sake. Applications are opportunity driven, but not the goal of the research. Hence scientific quality in the eyes of the individual researcher is independent of its potential to lead to applications for industry. Societal relevance however, in the eyes of an astronomer, arises self-evidently from the fundamental questions Astronomy gives answers to.

The last quote hints at another aspect of scientific quality, which follows from the astronomers' demand of good correspondence between discovery and reality: using sound scientific methods.

> "I guess [good quality research is] if you followed the methods as best as you can – like to the best ability and take everything into account and thoroughly test your results and outcomes to make sure that they are as concrete and solid as they can be before even throwing them out to the general populous … Part of it also is, if you have high quality data, it can be easier to do high quality research, so erm, that too." (Master Student 2)

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Hence, for good quality research “important questions” need to be answered by robust and careful research. This involves thorough methods, which ideally take all possible factors, assumptions and biases into account and sufficient testing of the methods and results before publishing. However, those criteria are yet not enough to satisfy an astronomer’s account for high quality research: Conclusions that push knowledge forward must not only describe ‘reality’ but must also be “rememberable” (Postdoc 2) and communicated well:

> “And you have written a paper which demonstrates you have answered that question […] And you have written it in such way that a non-expert in that field can read it and understand what you have done. They may not understand the details, they may not understand the algorithms, but I think high quality research is: You can pick up – a good paper – any Astronomy paper, read the abstract, read the introduction, read the conclusions and know what they did. And why they cared. And you may not know the shear statistics of galaxies of redshifts 2, but good quality research will give you the background and give the context which you should be able to understand. As a scientist you understand it. If it’s a crap written paper, then that’s crap research – I don’t care how brilliant the answer is, if they can’t communicate it through a paper or through a presentation, then that’s bad research. […] Yeah, I’d say that means high quality. They are able to write and present a compelling scientific argument from start to finish, that any reasonably trained human being can read and think about, you know.” (Faculty Member 4)

In summary, the study found that the astronomer’s definition of high quality research is based on three criteria:

1. Asking an important question for the sake of understanding the universe better and to push knowledge forward.

2. Using clear, verifiable and sound methodology.

3. Clear communication of the results in order for the community to make use of them.

As obvious as that definition may seem to an astronomer, interviewees admit that there is no easy answer to the question how it can be measured, whether those quality criteria are fulfilled. After all, “scientific quality is hard to measure, and numbers are easy to look at” (Benedictus & Miedema, 2016). That is why indicators serve as proxies to evaluate scientific quality and performance. In order to compare criteria applied in evaluation

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

procedures with astronomers' intrinsic motivation and what they value as quality research, the study investigated what the funding, publication and career systems value and which indicators are used, according to the astronomers. These findings will be explicated in the following section.

3.2 The Evaluation Gap in Astronomy

The former section presented the study's results of what an astronomer values in research quality. Now we will look at what astronomers perceive that is valued in the evaluation system. Only when we know more about the extent of the overlap we can investigate what effects indicator use in evaluation has on knowledge production and scientific quality, which is the aim of this study. We will start this section by giving insights into what our document analysis reveals about the Sterrewacht's strategies to maintain its success, since they likely affect targets that Leiden astronomers need to strive for. Leiden Observatory is ranked among the best Astronomy institutes in the world (Evaluation protocol[2010-2015]). The Sterrewacht is keen on maintaining that status by following NOVA's objective to "*ensure a front-line role in the next generation of astronomical discoveries*". NOVA intends to fulfil this objective by following its mission, which is to "*carry-out front-line astronomical research, to train young astronomers at the highest international levels, and to share discoveries with society.*" Leiden Observatory has three missions, which are well-aligned with NOVA's overarching one:

1. The Sterrewacht's educational mission is to "*to provide excellent education at the bachelor and master level, not only to prepare students for PhD projects, but also for the general job market.*"

2. Research at the forefront of modern Astronomy, including collaborations with Dutch partners such as TNO Delft, Dutch Space and the Sterrewacht's vicinity to ESA's ESTEC (Technical facility of the European Space Agency), enabling "*astronomers [to be] among the first to use the instrument, thus reaping the hottest early science harvest.*"

3. The Sterrewacht follows an outreach & education mission. Since "*Astronomy has a strong appeal to the general public*" (LU self-assessment[2010-2015]), all staff and students "*spend considerable time and effort to explain the exciting results of Astronomy to the general public, in the form of lectures, press releases and newspaper articles, courses, public days and tours at the old observatory complex, and input to television and radio programs.*"

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Institute evaluations take place every five years by an Evaluation Board (EB). The committee's review is part of the assessment system for all publicly funded Dutch research organizations, according to the Standard Evaluation Protocol (SEP). The SEP consists of three criteria: (i) Research quality, (ii) Societal relevance, and (iii) Viability. The scope of the assessment is set by the Terms of Reference (ToR), which in this case is the information provided by the self-assessment documents of the individual Astronomy institutes and NOVA as a whole. These documents are a description of the institute's mission, objectives and results. In addition, the EB conducts interviews with management, the research leaders, staff members, and PhD candidates.

In addition to the self-assessments prior institutional evaluations and those evaluations, Leiden Observatory measures its scientific productivity with certain "performance indicators" measuring the productivity of staff members and students: "*During the reporting period 2010-2015 Leiden Observatory thrived; its scientific production, measured in terms of number of papers, citations, PhD candidates and postdocs and the amount of grant money awarded, has never been so large*" (LU self-assessment$^{2010\text{-}2015}$). The Sterrewacht calls them "objective" as they are quantitative and they include:

• **Publications:** Total number of refereed papers.

• **Citation rates:** including 24 citation parameters (e.g. number of citations, number of normalised citations, number of normalised first author citations)

• **PhD theses**

• **External grants and prizes**

• **Outreach activities:** The performance of its outreach programme is measured by the large numbers of press releases, articles, attendees, teachers and children reached through its various activities.

• **International leadership:** International visibility of Leiden Astronomers and their leadership roles in organisations and committees.

• **Instrumentation programme:** A key indicator of the success here is the on-time, on-budget and within specification delivery of instrumentation (co-)built by NOVA. Another positive measure is the frequent invitations for international collaborations and the number of successful spin-off projects.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

In addition to the performance indicators, “excellence” is a rising buzzword to measure the success of institutes and researchers (Sørensen et al., 2015). The Sterrewacht “*believes that their success in winning international research funding demonstrates that their staff is of* high calibre *and has the drive and commitment to continue excelling. […] Their staff and the faculty board agree that excellence will be the most important hiring criterion.*”

The Sterrewacht commits itself to the missions and strategies described above, because the institute’s key goal is to “*maintain the present high level of achievement and to continue to score very well in international competitions for observing time at space observatories and on the ground, as well as for research grants.*”

The funding that the institute receives is comprised of baseline-funding from Leiden University and from NOVA. How much money the university allocates to each institute depends on a formula, which is called *“Allocatie Eerste Geldstroom”* (AEG):

> “So what we get from the university is determined by how well we have done over the last few years in terms of how many grants, how many PhD candidates, how much teaching we have done. It’s kind of an arrhythmic model that determines how much you get over the next year, it’s kind of a 3 year average.” (Faculty Member 2)

The amount of money that the university receives from the government is based on a similar formula. This model makes the institute very autonomous, but at the same time responsible for paying their staff. In addition to baseline-funding, the Sterrewacht vaunts the high number of external research grants acquired by individual staff members. The main funding agencies include the NWO and the EU European Research Council (ERC). The observatory reported already in 2009 (Annual report2009) that

> “*university funding is changing as a result of external pressures. There is more and more emphasis on temporary, project-based funding, threatening the structural long-term funding that is needed as the basis of a healthy scientific institute. Keeping up our success in funding applications is therefore vital.*”

This will become especially true during the next years when the continuation of the NOVA grant is running out in 2023 and NOVA needs to find a different source of funding. Outside grants are also “needed to fund graduate students” (Faculty Member 4) who are the ‘working horses’ of the system:

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> "[…] if you have money that means that you have also, that you have labour, that you have the effort available. And of course in exchange we need to define a little piece of science that that student can do as part of his PhD." (Faculty Member 1)

Grants are limited and very competitive. Money available for research is finite and so proposals need to fulfil certain criteria in order to be successful in acquiring grants. Advertising the so-called *"sexy topics"* are highly valued when the government and funding agencies decide which research proposal to fund. The interviewees frequently report that "the funding system is very much oriented towards the fashion of the day" as opposed to also "extremely important" topics that are "more pedestrian/ basic" (Faculty Member 1). Promising "impact" is important to acquire external funding and improving the AEG:

> "So the impact is very important, because if [the evaluation committee] had said that we are doing so-so or it is a field in decline or an institute that are not doing things right the university can start reallocating their priorities [as in funding]." (Faculty Member 2)

The potential impact of a research project is often estimated on basis of the recognition a scientist has gained, due to the common assumption that past achievements determine future outcomes (e.g. Kurtz & Henneken, 2017; Merton, 1968 & 1988). Achievement is measured in terms of quantitative indicators, such as publication- & citation rates, impact factors and the number of acquired grants or other "performance indicators" as described above. Given that such achievements determine an Astronomer's recognition and that recognition is needed to acquire more funding, we observe a *Matthew effect* in in Astronomy. This effect, the "Chicken-or-egg problem" (Faculty Member 1), means that past output determines future success (Merton, 1968 & 1988). The prevalence of the Matthew effect is frequently reported by interviewees, for example:

> "The funding system is mostly … Mostly looks at your *past achievements*, right? So much of what determines whether you get your next grant is what you did with the previous one, so … how that's evaluated or viewed, or judged, or measured is key." (Faculty Member 2)

> "Erm … the funding agencies have a tendency of – where the money follows the reputation. And strangely enough, it's not totally inappropriate that money follows

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> reputation. Erm, it is in a sense *Darwinian*, I mean something has success and therefore you should feed it, you should support it." (Faculty Member 3)

While serving the "fashion of the day" and promising impact seem to be the basis criteria for a successful grant application, the selecting process isn't very transparent and often depends on luck. According to Faculty Member 3, a career in Astronomy is "90% luck and 10% hard work" and this is partly because receiving funding depends on chance.

> "But … I think the biggest problem of the funding system how it is now, is that there is so little money available, that the selection problem is … I would say almost random, not completely random, but you could have a very good project and very robust project but not been given the money, because there are just too many." (Postdoc 2)

> "So you are good enough, that you know it's a good proposal. And you are now rolling the dice. You are just waiting for …. It will come down to: One person didn't have their chocolate biscuit in the morning and they are grumpy and they dinged you for not being concise enough." (Faculty Member 4)

This randomness in allocation of funding is what astronomers call the "TAC-Shot-Noise" (Faculty Member 4), which stands for "Time Allocation Committee"-Shot-Noise. The word 'time' here instead of 'funding' indicates that committees that grant observing time base their decisions on the same criteria as funding agencies. The interviewees make no difference between grants in terms of observing time and research money when talking about the funding system. In Astrophysics, having been granted observing time is generally as prestigious and important for one's career as funding. That grants are often based on luck generates a lot of (psychological) stress for applicants. Other consequences of this "rolling the dice" technique include tense competition and risk aversion to not lose out on impact. These (constitutive) effects of indicator use will be discussed in the next section. Despite the luck aspect, prestige and reputation are vital for receiving grants/ telescope time and career advancement due to the Matthew effect. The most prominent form of output are *first-author publications*, which are the capital of every astronomer. When an interviewee talks about "having a paper" or "publishing a paper" it is generally implied that that person is first author on that paper. Interviewees emphasise not only the importance of publishing, but also the "emergency" to do so, which results from a pressure to publish, another (constitutive) effect of indicator use, discussed in the next section.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> "Before you have a tenure job, you've got to make an impression and demonstrate that you can produce papers in a reasonably rapid fashion." (Faculty Member 4)

> "You always want to be the fastest and want to have your results out. But it's not really a deadline, it's more an *emergency*." (Postdoc 2)

While some astronomers claim that "everything that is not obviously wrong is publishable" (Faculty Member 1) other interviewees relativize this: "It's not sufficient to be true. It has to be true and pushing knowledge" (Faculty Member 4). This still matches with the astronomer's values, as generating output in the form of disseminating knowledge, including informing the public, is important for an astronomer. As elaborated in the former section, discoveries matter for their own sake and resulting applications are merely a bonus. However, the EB criticizes that "*valorisation seems to be opportunity driven, rather than to derive from pre-determined strategy.*" Hence, while the evaluation system demands for more directly applicable output to demonstrate society relevance of the research, astronomers do not intrinsically strive for such output.

Furthermore, what fulfils that criterion of 'being publishable' is often open to interpretation, so lies in the eye of the reviewer. Often it also depends on the research field. In the field of exoplanets, a detection with the right method (e.g. direct imaging as opposed to radial velocity) can be enough to publish already without interpretation or analysis. In the field of Radio-Astronomy that is the same case, as detections through long wavelengths are extremely difficult. Hence, in some observational Astronomy fields a *sole detection* is highly valued by journals and reviewers.

However, in observational Astronomy, *non-detections* are much more frequent than detections and about 90% cannot get published (Postdoc 1). Unless the non-detection can 'add to new knowledge' by having been able to calculate upper limits or demonstrate anomalies, they are not publishable:

> "[Negative results are not publishable], unless you have a very good, as in for example the way we sort of explained the upper limits with the non-detection. […] *The problem is how to tailor it, right?* […] So, yeah, unless you have … like a good way, I mean there is some research that published non-detection – for exoplanets sometimes they publish it when they didn't detect it, because sometimes you sort of predict that it should be there … […] And it's an anomaly or something like that …

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> […] So there are some ways to publish this, but I think it's very … like 10%. There is a whole 90% that doesn't get published and sometimes, like for example, if you just had bad weather, then it's very difficult, right?" (Postdoc 1)

In summary, to survive the climb up the career ladder, an astronomer has to acquire recognition on basis of quantitative indicators and publish enough first author papers. The ranking of the universities of previous job positions influences further career development. The Matthew effect leads to a "Golden Child Trajectory" (Faculty Member 4), where the 'ideal' career in Astronomy is a straightforward climb of the tenure track. This often involves committing to a professional life in the hamster wheel of the "cycle of observing, analysis and publishing" (Faculty Member 1).

From this investigation of values two opposite notions of science emerge. The first is the astronomer's "ideal" image of science (e.g. Postdoc 2), where astronomers are driven by their curiosity and the search for truth, limited only by epistemic restrictions such as technical possibilities, which was described in *Section 3.1*. The other notion is the image of a "system" of science, constituted by evaluation practices, such as indicator use, with values that are not in line with the astronomer's intrinsic values.

> "You [wouldn't be] bothered with the raw numbers. I have [*number taken out in order to assure anonymity*] refereed publications and [this] would probably be a lot smaller if your publication rate wasn't so important. […] Yeah, I have my doubts about the usefulness of that *system*." And: "Well, once again, I am not that happy with that cycle. It can put a lot of pressure. And I am trying to ignore that pressure now. I mean I am [above 50; *number taken out in order to assure anonymity*], so my career is established, let's put it that way, so I don't need to prove myself anymore, so I can safely ignore that pressure. But I think that younger people who still have to make that career have to work according to that *system* and I am not quite sure that that is actually a good thing." (Faculty Member 1)

> "The problem is – this is the main thing, right – if you wanna have a job later on, you are gonna a have to have papers, because that's how it works. Even though I don't like the *system*, I don't like the way it is, it is what it is and you have to adapt to it." And: "It's the same issue, it's the same thing … It's a *system* problem I think. Erm, I try to do quality research, but I do feel sometimes that I end up publishing because I have to publish." (Postdoc 1)

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

The discrepancy between those two notions of science gives rise to the evaluation gap. Therefore, we can say that constitutive effects of indicator use generate the evaluation gap; indicator use leads to a concept of quality which is not the same as astronomers would define it. In turn, the presence of the evaluation gap has shaping consequences on the research behaviour and knowledge production in Astronomy. Fig. 20 illustrates the evaluation gap and its constitutive effects on motivation and identity. Those effects are outlined in the following section.

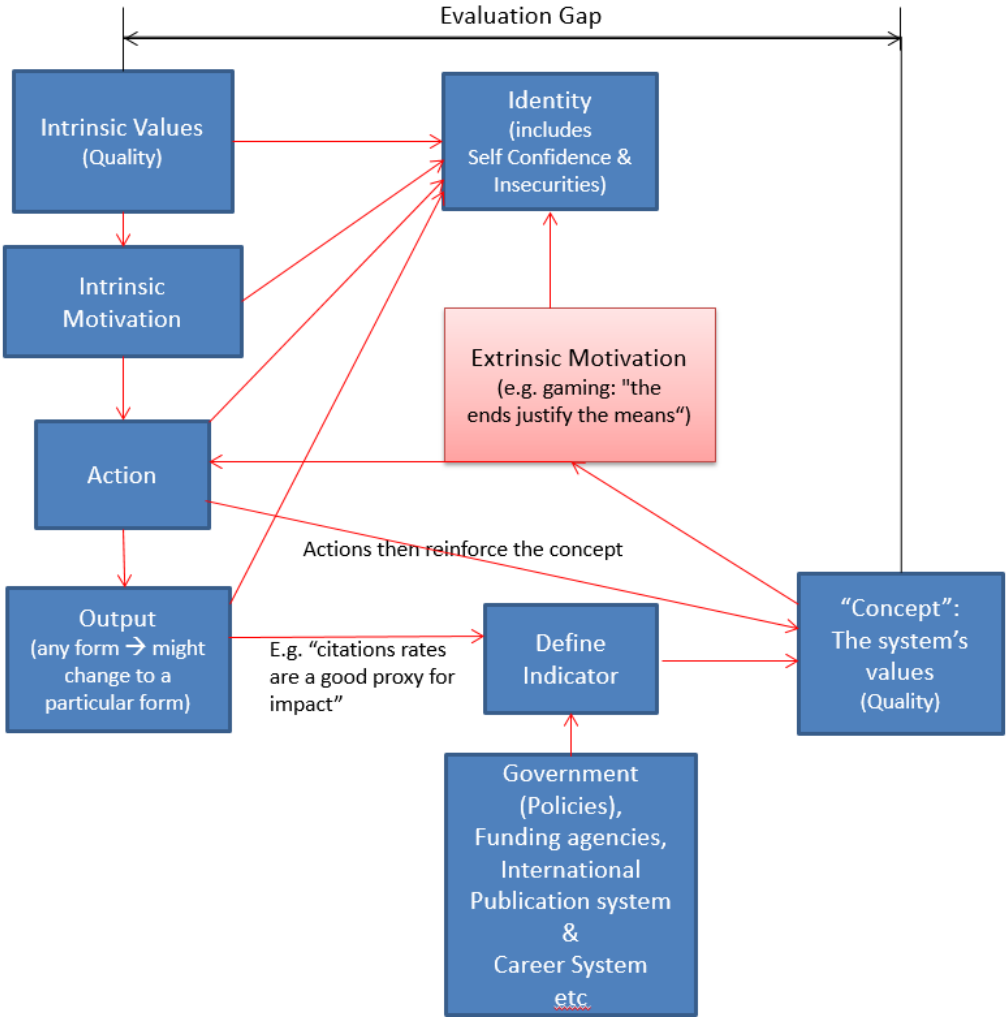

Fig. 20: *This figure illustrates the evaluation gap and constitutive effects in Astronomy. The cycle starts with the astronomer's intrinsic values and shows what constitutive effects (red arrows) each element has. Because the system does not have constitutive effects on the astronomer's intrinsic values, it also does not influence their intrinsic motivation. The evaluation gap between what an astronomer values and what is actually measured has constitutive effects on the astronomer's identity in form of psychological effects, for example feelings of unworthiness, as outlined in the next section.*

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

3.3 Constitutive effects of indicators on knowledge production in Astronomy

We have found that performance indicators in research in Astronomy do not reflect the astronomer's definition of research quality. That gives rise to an evaluation gap, which, as we found, have consequences (i.e. "formative effects"; Dahler-Larsen, 2014) on research behaviour and knowledge production. Because those consequences are formative, they are *constitutive in their effects*. The indicator then stands in "a constitutive relation to the reality it seeks to describe" (Dahler-Larsen, 2014). Meaning is being constructed (e.g. citation rates equals impact) and practices are being established ("pushing publications"; e.g. Postdoc 1 & Faculty Member 4). Without being exhaustive, Dahler-Larsen (2014) distinguishes between five main categories of constitutive effects of indicators: indicators define interpretive frames and world views (*A*), social relations and identities (*B*), content (*C*), time frames (*D*) and change in their meaning as a consequence of their use (*E*). This section will portray constitutive effects arising from the evaluation gap in Astronomy and relate them to one or more of those categories (performed by top-down coding).

According to this study's interviewees, output in the form of papers and its quality assessment through quantitative indicators such as publication & citation rates, defines the value of an astronomer (*B*). The increasingly limited number of jobs the higher up the career ladder, introduces a highly competitive "rat race" and "postdoc circus" (Faculty Member 2 & Faculty Member 1). Astronomers need to need to acquire recognition in the form of quantitatively measurable output to establish themselves in the community.

> "It's just because there is so much competition, that the first filter you go into is how many papers you have. Doesn't matter how good or bad «laughing» they don't check this that much." (Postdoc1)

The need for this kind of output to survive "filters" on the career ladder ("publish-or-perish" e.g. Master Student 2 & Faculty Member1), however, causes pressure (*D*).

> "And I don't know … I think also if there was less pressure … Financial pressure to conduct research, people would not have to resort to stupid tricks. And trying to make themselves appear more … high quality researchers than they are by for example publish too many papers or publishing wrong things or hasty or too

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

> quickly without taking too much care. I think indeed, the lack of funding is … is hurting the research quality. Not really in the sense that we don't have enough money to do all the research that we want, but it's affecting the research that is being done, by *sacrificing quality for efficiency*." (Postdoc 2)

Publication pressure is always "at the back" of an astronomer's mind (Postdoc 1) and the pressure may increase when one or more of the following factors are present, because they define the time-frame of publishing (*D*):

• First, when the astronomer faces head-on competition, there is a *race for priority*, which pushes the researcher to "publish as fast as I can … as soon as I get the data" (Postdoc 1).

• Second, timescales of projects and publishing are "tied to the *timescale of [PhD] students and postdocs*" (Faculty Member 2), because "they need to get their thesis chapters out. They need to be ready for the job application season".

• Third, *telescope application deadlines* are perceived as "natural" deadlines (e.g. Faculty Member 2) for publications, as performance indicators such as the publication rate are part of the assessment criteria for observation time allocation.

Publication pressure may have psychological effects, such as demotivation, discouragement and feelings of unworthiness, on the researcher (*B*) and constitutive effects on the (quality of the) content (*C*). The latter may include cutting up publications in order to publish more ('salami slicing'), premature publishing, and non-replicable papers (*C*).

> "From looking at people who are doing PhDs, erm, you know there is still, they are in on weekends, they are doing more than 8 hour days, they are doing more than 40 hour weeks. You know, they don't take the full amount of holidays allocated to them, which I didn't realise. […] As much as I have been told, that this university really encourages you to have a life outside your PhD, I see very few examples of that. And the examples of that, that I see, are people who […] basically don't let themselves be *bullied by their supervisor* into feeling that they have to do all of this additional work. Some people are happy with this, but I don't want my entire life to be one thing …because it causes me *too much stress for my entire life to be in academia*. […] I think I figured out that it would be *constantly proving that I was good enough*. Constantly proving that I was worth the money, constantly proving that, you know, I

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

was worth the time and the energy and all of that and that sounded exhausting before I even started it. And sounded like I would constantly be battling with feeling I am not good enough, while trying to tell other people that I was good enough. And I kind of went 'No' – and I am not – I know, there is gonna be an element of that in jobs as well, but I feel a little bit, in a job, at least there should be a break, like this is 9-to-5 or whatever. And then I can go home and I can leave it there. Whereas with academia, it's kind of like, yeah you can go home, but then you are getting emails, until maybe 20:00 or 21:00 in the evenings and still doing things." (Master Student 2)

"I think sometimes yes, the pressure to publish has forced us to sometimes push out results, where having another observation or two would make a significant improvement on the current results." (Faculty Member 4)

"[…] they would skip some tests, obvious tests, that they could have done, but that maybe take a bit of time, or that they use a method without properly characterising the biases or the assumptions that are used behind this method. " (Postdoc 2)

"As an observational person you should be able to publish all your data reduction scripts from start to finish. It spits out the output files, which you see in the paper and then somebody else can come along. And I know the reason why is that: There is a fear that, because you made it easy for other people to check your code, other people can find your bugs more easily and so you may get criticized for having buggy code over somebody who never publishes their code and bugs are hidden for years and years and years. There is no incentive at the moment to publish the code."(Faculty Member 4)

"And to be fair, it's mainly because if you want to have a paper, it has to be something new. Sort of. So you are not going to be publishing, checking that someone else's work is fine. That's not gonna give you a paper. You have to either find that something is wrong on the paper or you have to find the same and something more, right? Like, adding to it. So I don't know how much gets checked. I don't think a lot. But I do think if you read a paper and try to reproduce it, it's not very easy from a paper." (Postdoc 1)

In most cases, according to the interviewees, those effects have a negative effect on research quality. Publication pressure however, can also have positive effects, focusing and confining the research question. Salami slicing can be beneficial for good communication and readability of research results. The interviewees, however, remarked that often results are difficult to replicate. This is because of the lack of incentives to

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

publish information needed for replication, such as code used for analysis and a lack of incentives for reviewers and no dedicated time frame for the reviewing process (*D*). Prematurely published information or insufficiently reviewed papers make output even less readable and reproducible, reducing content quality (*C*).

Quantitative indicators define the landscape of success and its inverse: the landscape of failure (*A*). The interviewees have a hard time defining 'failed research', due to the very risky nature of research. They are only confident to describe what bad research is – the opposite of good quality research according to their definition (i.e. the three criteria). In contrast, the community and the system do have a definition of failed research, viz. the opposite of successful research as measured by indicators. The use of indicators then causes a shift in what counts as new discoveries in the community, from "anything new" (e.g. Master Student 1 & Postdoc 1) to "publishable results" (e.g. Master Student 1, Faculty Member 2 & 4). Hence, as long as negative results (e.g. non-detections) can't be put in a context ("tailoring"; Postdoc 1) where they become publishable, they are regarded as worthless: the research project failed and the researcher feels like a failure (*A, B*). This is despite the fact that in many fields in Astronomy non-detections are far more common than detections. Because those are hardly made public, astronomers express their frustration with the 'wheel being reinvented' and hence resources wasted. Especially young researchers can't afford to take on too risky research projects, which causes risk aversion ('playing it safe'; Stephan, 2012) and a tendency to prefer sexy topics to equally important non-sexy ones. This again has effects on research agendas and content (*C*).

> "These young folks are scared! They are afraid! […] And you know what, that is ultimately bad. This is ultimately bad, because in such a science where you know, ignorance is so big, being scared is not the right thing to be. […] You get results by your brains, your hands, by the collaboration with your colleagues and stuff, but you have to have a sort of courage. And it is *bread out of the young people*. Because they are not rewarded for their courage. And I find that very, very, *very bothersome*. That generation – people growing up like that. How are you ever, ever, ever going to understand the universe if you don't have courage?" (Faculty Member 3)

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Thus, the evaluation system undermines astronomers' values by putting a (too) strong focus on quantitative indicators. As a consequence, an astronomer's motivation also shifts to an output orientation where safe and accessible projects become the driver. While "the publication is not the aim – [it] is a means to showing what your methodology is" (Faculty Member 2) it does become an aim. Producing high quality research "to know and understand better" and communicating this knowledge to the community is what makes up an astronomer's intrinsic motivation to conduct research. However, the need to survive the climb up the career ladder gives an extrinsic motivation to perform research, which is oriented towards hitting the required targets (*B, E*).

> "And ahh … if [publishing] was not so important [to keep your standing] … I mean I would still publish my papers [but] it *gives a different motivation* to it, right? As a scientist you just want to publish your papers, because you are a scientist and you think this is important for science: 'This is the result, this is what defines the process of science'."
>
> (Faculty Member 1)

The Sterrewacht as an elite institute is such a compelling case since its mission to maintain its success, which is largely measured by indicators (as listed above) provides the right conditions for an evaluation gap. A higher pressure to achieve targets may lead to astronomers adapting their definition of quality to what the evaluation system measures in order to survive in the system. That is why it is all the more interesting that, even under the conditions set by an elite institute, we found that the astronomer's definition of quality remains unchanged. In other words, the results also show that, while indicators give an extrinsic motivation to an astronomer to perform, their constitutive effects do not reach as far as to affect an astronomer's intrinsic values to a noteworthy extent (see Fig. 20). This is the reason why the "ideal" and the "system" accounts of science do not conflate and astronomers try to serve both – evaluation gap remains.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

3.4 The Balance Act – Reconciling the concepts 'Evaluation Gap' and 'Constitutive Effects'

In the previous sections, we outlined how the discrepancy between what astronomers value as scientific quality and what they perceive what indicators measure constitutes an evaluation gap. This evaluation gap, in turn, has constitutive effects on researcher's motivation and the knowledge production process, including the resulting research quality. However, since the interviewees try to serve both imperatives, the "ideal" and the "system" one, at the same time, we could not observe any substantial constitutive effects of indicator use on the astronomer's intrinsic values and motivation. As a consequence, the two notions do not conflate and indicators are not the only "way through which the world is defined" (Dahler-Larsen, 2014). This is the reason why, at least in the case of Astronomy, it makes sense to use both concepts, the 'evaluation gap' and 'constitutive effects' in order to reflect on the effects of indicator use.

One particular constitutive effect of the evaluation gap is the advent of a third notion of science: coping with the system. Astronomers try to manage a balancing act between their intrinsic values and the requirements of the system. According to Dahler-Larsen (2014) indicators "define a strategic landscape in which practitioners must navigate". In the case of an astronomer, the strategic landscape is situated between the astronomer's intrinsic values and those defined by evaluation practices and the Sterrewacht's missions. That is where the balancing act takes place.

In particular early career interviewees struggle with the balancing act between performing high quality research according to their standards and fulfilling the requirements of the system. Because success in science or in the scientific career is not only dependent on quantitative indicators, but also on luck (e.g. Faculty Member 3; "90% luck and 10% hard work"), especially young researchers have psychological struggles with this uncertainty. Van der Weijden (2017) elaborates on this further. For some researchers, this discrepancy is unacceptable. As a consequence, they wish to leave academia:

> "Yeah [I don't want to stay in academia], partly because there is this 'publish or perish' thing, where it seems to be like 'pump it out'." (Master Student 2)

However, astronomers may also accept "the system" as a "fact of life" (Waaijer et al., 2017) and decide to "deal with it". The third notion can be described as a synthesis or mix

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

between the other two notions – the “ideal” and the “system”. When astronomers master the balancing act between staying true to their own value’s, while at the same time fulfilling the quantitative requirements, when they are being practical with respect to their work, they find a middle ground where psychological struggles are minimised as the astronomer accepts “the system” and adapts to it. We can observe this in interviews, where especially tenured astronomers describe how they practically “deal with the system” (Faculty Member 1) in terms of getting funding, telescope time and publishing. They emphasise how their science is observation-driven (e.g. Faculty Member 1 & 4) and explain how artificial deadlines, such application deadlines for telescope time, are “natural” deadlines to them (Faculty Member 2). Because of managing the balancing act, tenured astronomers feel that their work is generally in line with their criteria of quality. While having to “adapt to the system” which they do not like (Postdoc 1), early career researchers also declare that they would not personally compromise on quality too much, because research quality “is more important than ultimately [their] career” (Postdoc 2).

Almost all interviewees – even those tenured astronomers, who feel that their research is in line with their notions of quality – acknowledge problems of the “publish-or-perish-system”. Master Student 2 observes that “people talk about the publish-or-perish thing and how it hurts. And then other people seem not to have much of an issue with it.” On the one hand, astronomers know they need to play along with the system. On the other hand, they know what “really matters” (Faculty Member 2).

> Annual report[2014]: “With 16 PhD theses and 318 refereed papers, the scientific 'production' was fantastic. However, in 2050 it will not be those kinds of facts that count, it will be the *true discoveries that have stood the [test] of time that will be remembered*.”

As we observed that in practice research quality is harmed in many respects, either the amount of astronomers who manage the balancing act without sacrificing research quality is extremely low, or there is a fine line between working according to the third notion of science and a bouncing between the “ideal” and the “system” notion of science, where quality is sacrificed at least occasionally and justified by having to survive in the system. More research will be done on this matter.

In any case, whether astronomers manage the balancing act and work according to this third notion of science, or they flip between the two other notions, the majority of astronomers seem to indeed accept the pressure to publish as a “fact of life”. Waaijer et al.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

(2017) find that being able to cope with the system enhances the early career researcher's sense of autonomy and independence. In addition to their intrinsic motivation, this is probably why so many early career interviewees state that they "try to stay in academia for long as possible" (PhD Candidate) and why pressures can even be party self-enforced (Waaijer et al., 2017):

> Postdoc 2: "If at all possible, yes, I would like to continue in academia. And in a way this rule I have – 1 paper per year – is the standard I have posed on myself in order to have a good chance to continue."

On the one hand, this would be consistent with Waaijer et al. (2017) who claim that, while many PhDs (from different fields) state that publication and grant pressure is too high and had made them hesitant to choose a career in academia, it has not been a decisive factor in their actual job choices. On the other hand, early career interviewees are aware of the fact they might have to leave academia and are working on accepting that. Thus, to what extent a third notion of science, can be held by astronomers in practice, and early career researchers in particular, is subject to future investigation. It would be interesting to see whether or not such a third notion implies a bias towards perceiving the positive aspects of "the system" in order to guarantee one's survival on the career ladder, which would give justification for sacrificing scientific quality. I suggest to employ the Rational Choice Theory to investigate the workings of the balance act and how it is related to individual situations, since that approach sheds light on the factors that go into decision-making processes of individuals. Those could be classified as different typologies of coping. The logic of aggregation will then show what the different coping strategies, as part of the balance act, mean for the quality of science.

4 Conclusion

We have analysed 9 interviews with astronomers from Leiden Observatory and a collection of (self-) evaluation documents and annual reports from that institute and the Dutch astronomy umbrella organisation NOVA. We have elaborated on what values drive an astronomer to enter academic research and how they perceive the values of the publication, funding and evaluation system. We then analysed how the astronomers' values relate to the system's values and what constitutive effects a discrepancy – the evaluation gap – has on knowledge production in Astronomy.

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

We found that astronomers are driven by curiosity, truth-finding and “pushing knowledge forward”. During discussions of the interviews with CWTS's group for Science and Evaluation Studies[1], the question was raised of whether these values are based on a folk theory based on the public’s *enchanted view*[2] of how science works. A folk theory is a belief based on received wisdom, rather than concrete evidence and facts. However, while especially young astronomers are likely to hold an enchanted view about science and may become disillusioned by their experience in academia (e.g. Postdoc 2), we have observed that the astronomer's intrinsic values hardly change due to this disillusion. Therefore, we conclude that the astronomer's values are based on the realist account that astronomers generally hold, rather than on a folk theory about scientific quality. Astronomers derive scientific quality from their values, and define quality as ‘objective’ when it meets those values. We found that the astronomers’ account for scientific quality is based on three criteria:

Quality-Criterion 1: Asking an important question for the sake of understanding better and to push knowledge forward.

Quality-Criterion 2: Clear, verifiable and sound methodology.

Quality-Criterion 3: Clear communication of the results in order for the community to make use of them.

While astronomers agree on what quality is, they do admit that it is difficult to measure. Because resources such as funding and positions are limited, proxies for scientific quality – the quantitative indicators – help decide whom or what to fund. Those indicators include bibliometric measures such as H-indices, citation and publication rates. They also include the amount of funding acquired and how much observation time an astronomer has been granted. The more prestigious the affiliations a researcher had, the better their profile and chance to climb up the career ladder.

In order to survive in the current science evaluation system, which includes the funding, publication and assessment systems, the astronomer needs to fulfil the requirements of what is valued in “the system”, as constituted by quantitative indicators. The discrepancy

1. *https://www.cwts.nl/research/research-groups/science-and-evaluation-studies*

2. Science in Transition, Position Paper 2013, *http://www.scienceintransition.nl/*

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

between the astronomer's and the system's values gives rise to an evaluation gap (**Fig. 20**). We found that the evaluation gap in turn has a variety of constitutive effects on knowledge production, ranging from research agendas, researcher's behaviour and identities to research content.

There is a shift of focus from high quality science to publishing a high number of papers, presenting research that is less robust, replicable, and transparent than aspired. Risk aversion discourages creativity in the scientific process which inhibits innovative ideas, while valorisation gains ever-growing importance.

Interestingly, we observed that the astronomer holds two opposing notions of science: the "ideal" one which corresponds to their intrinsic values and the "system" notion. This means that, while in their daily research life an astronomer adopts an extrinsic motivation to perform science, their intrinsic values and motivations remain as their ideals. Hence, while indicators give an extrinsic motivation to an astronomer to perform, their constitutive effects do not reach as far as to affect the astronomer's intrinsic values to a noteworthy extent. However, constitutive effects of indicator use may not shape the realist's notion of reality, but they do shape research agendas and have epistemic implications on day-to-day research practices. Man-made deadlines become "natural" deadlines. As a consequence, the evaluation gap remains and a third notion of science arises: coping with the system. The astronomer always tries to manage the balancing act between their intrinsic values and the requirements of the system. In order to do so, astronomers must accept "the system" as a "fact of life" Waaijer et al. (2017), serving to quantitative indicators, while at the same time not sacrificing research quality. Unsurprisingly, we found a difference between early career researchers and established faculty members, where the former struggle with the uncertainty ahead, often considering to leave academia and the latter being more confident that they are managing the balance act. However, providing a typology and different 'coping strategies' is subject to further research.

We conclude that Leiden observatory's goal of "*fostering an intellectually rich and vibrant scientific atmosphere which allows astronomers to pursue their ideas and push scientific boundaries, and in which young scientists can develop and grow*" is not compatible with its strategy to ensure a front-line role in Astronomy if this front-line is defined by quantitative indicators. Instead, we propose to find alternative indicators, whose constitutive effects could be utilized such that the evaluation gap is minimized. By means of "innovative use"

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

of indicators (Fochler & De Rijcke, 2017), positive constitutive effects could alter researchers' behaviour and to regulate the knowledge production process to privilege scientific quality. In such a scenario, the institute's goal could be met as the astronomers could act upon their intrinsic motivation, while at the same time being extrinsically motivated to perform at a high level. While there is currently little literature on the topic, future investigations into alternative evaluation practices (Duffy, 2017) and innovative indicators ("*re-configuring evaluation*"; Fochler & De Rijcke, 2017) have been proposed. As this study demonstrated, for this kind of future research it makes sense to consider using both concepts, the 'evaluation gap' and 'constitutive' effects in order to describe the reactivity of indicator use.

References

Annual report1998- Annual report2015: Annual public reports, authored by the director of the institute (1998-2015), https://www.strw.leidenuniv.nl/research/annualreport.php

Austin, J.L. [1962] 1975, "How to Do Things with Words", [Cambridge, Mass.], Harvard University Press

Benedictus, R. & Miedema, F. (2016), "Fewer Numbers, Better Science", Nature, Vol 538, p.453-455

Bourdieu, P. (2004), "Science of Science and Reflexivity", The University of Chicago Press, ISBN: 9780226067377 & ISBN: 9780226067384

Bouter, L.M. (2008), "Knowledge as Public Property: The Societal Relevance of Scientific Research", OECD, http://www.oecd.org/site/eduimhe08/41203349.pdf

Callon, M. (ed. 1998), "The laws of the market", Oxford: Blackwell

Dahler-Larsen, P. (2014), "Constitutive Effects of Performance Indicators: Getting beyond unintended consequences", Public Management Review, 16:7, p.969-986

De Rijcke, S., Wouters, P., Rushforth, A., Franssen, T. & Hammarfelt, B. (2015), "Evaluation practices and effects of indicator use – a literature review", Research Evaluation, 25(2), p.161–169

Desrosières, A. (1998), "The Politics of Large Numbers – A History of Statistical Reasoning", Harvard University Press, ISBN 9780674009691

Duffy, D. (2017), http://blogs.lse.ac.uk/impactofsocialsciences/2017/08/14/rather-than-promoting-economic-value-evaluation-can-be-reclaimed-by-universities-to-combat-its-misuse-and-negative-impacts/

Espeland, W.N. & Stevens, M.L. (2008), "A Sociology of Quantification", European Journal of Sociology, Volume 49, Issue 03, p. 401 – 436, https://doi.org/10.1017/S0003975609000150

Espeland, W.N. & Vannebo B. (2008), "Accountability, Quantification, and Law", Annual Review of Law and Social Science 3, p. 21-43

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Fochler, M. & De Rijcke, S. (2017), “Implicated in the Indicator Game? An Experimental Debate”, Engaging Science, Technology, and Society 3, p.21-40

Fochler, M., Felt, U. & Müller, R. (2016), “Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists’ Work and Lives”, Minerva, Doi: 10.1007/s11024-016-9292-y

Foucault, M. (1977), “Discipline and Punish: The Birth of the Prison”, London, Allen Lane

Foucault, M. (2003), “The Subject and Power”, in Rabinow Paul and Nicholas Rose, eds., “The Essential Foucault” (New York, The New Press, p. 129-144)

Godin, B. (2006), “On the origins of bibliometrics”, Scientometrics, Vol. 68, No. 1, p.109-133

Hammarfelt, B. & De Rijcke, S. (2014), “Accountability in context: effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University”, Research Evaluation (2014), pp. 1–15, doi:10.1093/reseval/rvu029

Heidler R. (2011), “Cognitive and Social Structure of the Elite Collaboration Network of Astrophysics: A Case Study on Shifting Network Structures”, Minerva (2011) 49:461–488, https://doi.org/10.1007/s11024-011-9184-0

Kaltenbrunner, W. & De Rijcke, S. (2016), “Quantifying ‘Output‘ for Evaluation: Administrative Knowledge Politics and Changing Epistemic Cultures in Dutch Law Facilities”, Science and the Public Policy, p.1-10

Kurtz, M. J. & Henneken, E.A. (2017), “Measuring Metrics - A 40-Year Longitudinal Cross-Validation of Citations, Downloads, and Peer Review in Astrophysics”, Journal of the association for information science and technology, 68(3):695–708

Laudel, G. & Gläser, J. (2014), “Beyond breakthrough research: Epistemic properties of research and their consequences for research funding”, Research Policy 43, p.1204-1216

LU self-assessment2010-2015: Self-evaluation Document University Leiden, Leiden Observatory (2016). Period 2010-2015

MacKenzie, D. (2006), “An engine, not a camera. How financial models shape markets”, Cambridge, MA: MIT Press

Merton, R. (1936), “The unanticipated consequences of purposive social action”, American Sociological Review, Vol. 1, No. 6 (Dec., 1936), p. 894-904

Merton, R. K. (1968), “The Matthew effect in science”. Science 159: p.56–63. Page references are to the version reprinted in Merton (1973).

Merton, R. K. (1988), “The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property”. Isis 79: p.607–623.

NOVA self-assessment2010-2015: Self-evaluation Document NOVA (2016), including NOVA’s evaluation of its four institutes and a joint appendix. Period 2010-2015

Porter, T. (1995), “Trust in numbers”, Princeton University Press

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

Rosanvallon, P. (2009), “Demokratin Som Problem”, Hägersten: Tankekraft Förlag.

Rosenberg, N. & Nelson, R. (1994), “American Universities and technical advance in industry”, Research Policy 32, p.323-348

Rushforth, A.D. & De Rijcke, S. (2015). “Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands”, Minerva 53, p.117-139

Sørensen, M.P., Bloch, C. & Young, M.(2015), “Excellence in the knowledge-based economy: from scientific to research excellence”, European Journal of Higher Education, https://doi.org/10.1080/21568235.2015.1015106

Stephan, P. (2012), “How economics shapes science”, Harvard University Press

Van der Weijden, I.C.M., Meijer, I., Van der Ven, I., Ali, R.F. & De Gelder, E. (2017), “The Mental Well-Being of Leiden University PhD Candidates”, https://www.cwts.nl/download/f-x2q213.pdf

Vestman, O. K. & Conner, R. F. (2006) “The Relationship Between Evaluation and Politics” in Shaw, I. F., Greene, J. C. and Mark, M. M. (eds), The Sage Handbook of Evaluation 225–42. New York: Sage Publications., in Dahler-Larsen, P. (2014)

Waaijer, C.J.F., Teelken, C., Wouters, P.F. & Weijden, I.C.M. (2017), “Competition in Science: Links Between Publication Pressure, Grant Pressure and the Academic Job Market”, High Education Policy, https://doi.org/10.1057/s41307-017-0051-y

Supplementary Material

1. Interview Questions:

Topic	**Research Question**	**Interview Question**	**Type Respondent (Faculty, Postdoc, PhD, Master)**
Introduction	Background	E.g. How did you get this position, which career steps were necessary?	All
	Topic (How much does the choice of the research topic depend on the need to get funding? (avoiding risk taking?))	What is the topic of your research?	All
Project funding	Conditions of funding	How did you received funding for this project?	All
	Institutional conditions of funding	How is funding allocated in your institute in general?	All
Exposure to assessments	What role do assessments play in an astronomer's (daily) life?	What role do assessments play in your work?	All
		> Do you have yearly appraisals/ R&O talks with your supervisor? Peer review for funding applications & mid-term reviews for projects?	All
		Are you held accountable to the founder/ review panels on a regular basis?	All
Knowledge production – Planning research	What is the choice of topic dependent on (e.g. preference of supervisor/ funding/ own interest/ riskiness)?	How do you decide on a topic for your research?	All

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

		What do you advise PhD students when they ask about how to select a research topic?	Senior – Faculty
		How do you give priority on topics if you have more than one to work on?	All
		Is the journal agreed upon before writing? So, does the choice of the research topic, methodologies and content of the paper depend on that choice?	All
Knowledge production - Doing research	What are the effects on choices about the research process? (e.g. Effect on methodologies used?)	What needs to be taken into consideration for designing a Methodologies Project Design?	All
		Do you feel restricted in the research process?	All
		Have you heard about "responsible research methods"? And what's your stand towards it?	All
	Does the evaluation system foster collaboration or lead to competition?	How is collaboration organised in your project/institute/field?	All
Knowledge production - Publishing research	Is publication pressure a result of the evaluation system? And how does it influence the publications (e.g. premature publishing/ salami slicing)?	What are the most important factors in your field for deciding on when to publish research results?	All
		What are the most important factors in your field for deciding on what to publish [sexy results etc]?	All
		Do you perceive publication pressure?	All
		> Have you observed that people publish before the research has reached a more matured stage?	All
		> Have you observed that people cut up your research just to produce more papers of it?	All
	Does the evaluation system influence content?	Do you feel like you need to concentrate more on quantity than quality of your work?	All
		> Would one write up results differently if it weren't for the specific requirements measured by indicators such as impact factors and citation rates?	All
	How to deal with unexpected outcomes and "failures"?	What do you define as 'failed' research?	All
		Have research lines you have been engaged in ever failed?	All
		> If yes, what were the consequences in terms of funding, publishing etc?	All
		> If no, do you sometimes worry about not delivering the expected outcome due to a threat of not receiving further funds?	All
		Do you report "negative results"? Can they be published? Do astronomers/ you think that they should be published?	All

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

	How does the evaluation system influence replicability?	Do astronomers try to ensure that their published data is replicable or do you feel the necessity to keep information closed off?	All
What is quality in astronomy (value, quality, excellence)	Field: What is quality research in the field?	What is high quality research in your field?	All
	Institute: What is quality research in the institute?	How is high quality research defined in your institute?	All
	Researcher: What is quality research for the individual researcher?	What does high quality research mean to you?	All
	Researcher: What are motivational factors?	What drives you in your research?	All
	How does the funding system relate to good science quality as defined by the astronomer?	Does the funding system encourage good science?	All
	How does the publication system relate to publication quality as defined by the astronomer?	How does the publication system reflect upon quality in science?	All
		(Is the quantity of publications put above quality?)	All
Improving research evaluation & Consciousness	Are there wishes/ways to improve the evaluation system?	What issues do you think need to be improved to guarantee better science?	All
	Consciousness about the evaluation system	Do you feel that you are given the chance to question how science is performed?	All
	How did the system change over time and what did senior researchers observe?	When did you have your first encounter with the way science is performed and assessed? How did that compare with your initial motivation to become a scientist?	Senior – Faculty
		In your experience, did the definitions of value and academic quality change over time?	Senior – Faculty
	Do young researchers perceive that they need to adapt to the evaluation system?	When did you have your first encounter with the way science is performed and rewarded? How did that compare with your initial motivation to become a scientist?	Junior – Faculty, Postdoc, PhD, Master
		Can you pick topics and methods yourself or do you feel like you'll only be free to do that once you reached tenure?	Junior – Faculty, Postdoc, PhD, Master

Julia HEURITSCH
DOI: 10.3217/978-3-85125-668-0-09

2. These following codes represent themes which emerged by combining sensitivity towards existing literature on constitutive effects of indicator use with insights from our data. The interviews were coded using these codes:

Code	Explanation & Related Keywords
CAREER Clarity/ Expectations	Has the path been clear? What is expected in terms of career steps? Tenure.
Politics	
Prestige	
Output orientation	Both, in terms of output = basis of assessment & what output is expected.
Pressure	Publication/ Funding
Impact	
Competition	
Collaboration	
Riskiness	
Failure	
Negative results	Non-detections
Authorship	
Salami slicing	
Quality	
Curiosity	“Wanting to understand”
Referees	
Matthew effect	
Citation rates	
Publication rates	
Funding	
Gaming	Strategies, Targeting, "Sales men"
Replicability	
Epistemic Subculture	Topic of research, Instrumentation/ Observational/ Theoretician
Sexy topics	
Uncertainty (research)	
Uncertainty (career)	
Integrity	Fraud, Fake, Cheat
Luck	
Indicator	