Identifying Research Quality in the Social Sciences

Michael Ochsner

ETH Zurich, Zurich, Switzerland, and FORS, Lausanne, Switzerland

ochsner@evalhum.eu

https://orcid.org/0000-0002-1758-9590

Chapter accepted for publication in Engels, T. C. E., & Kulczycki, E. (Eds.). (2021). *Handbook on Research Assessment in the Social Sciences*. Edward Elgar, ISBN 9781800372542.

Abstract

What is good research? A seemingly simple question reveals itself difficult to answer. The current methods for identifying research quality come with validity issues due to a lack of conceptual scrutiny. I argue that research quality is a context dependent latent construct. While identifying research quality is a difficult task, it is not impossible. The methodological toolbox of the social sciences provides instruments to capture such latent constructs. I propose a method to conceptualise research quality in its context and argue that it is fruitful to combine peer review and indicator-based evaluations rather than playing them off against each other. Similarly, instead of juxtaposing notions of quality of different stakeholders, it is more promising to start from the scholars' notions of quality and to add other stakeholders' notions of quality in a communicative process arriving at a context-specific definition of research quality.

Keywords: Research Quality; Stakeholders; Latent Construct; Validity; Informed Peer Review; Measurement

Introduction

What is good research in your discipline? A simple question that should not be difficult to answer for any scholar. It is the basis of all research and scholarship. If one cannot identify good research, how can one possibly conduct high quality research? Still, this question poses problems, especially for scholars in the disciplines attributed to the social sciences and the humanities (SSH). In the natural sciences, very often, this question does not startle the academic community as much as in the SSH. Their answer would be: a piece of research that has a lot of citations in the relevant international databases. A current and extreme example is the answer by the French epidemiologist Raoult when confronted with critique on his research practice. He simply states that he and his team is the most highly ranked research group in the field by a webservice that ranks clinicians according to the number of publications and citations, which he sees as the prove that his research is the highest quality and thus accusations of fraud from scholars with a lower reputation cannot be of any meaning[1].

The other possibility is the claim that experienced researchers are qualified to make a holistic judgement: As an expert, I am able to recognise good research. By being capable of doing good research, a scholar is capable of making judgements on the quality of other's research without providing detailed arguments. As an interviewee in a study on peer review processes put it: "Academic excellence? I know it when I see it" (Lamont, 2009, p. 107; see also Gozlan, 2016; Johnston, 2008; Van den Brink & Benschop, 2012). While this possibility is documented for all disciplines, it is the primary method used by SSH scholars – at least by those who do not argue that it is impossible to qualify research.

[1] See the interview on his youtube channel: https://www.youtube.com/watch?v=mJl2nPHAo2g&feature=youtu.be

Not surprisingly, these two ways of judging the quality of research are the two main methods used in formal research evaluations as well: Indicator-driven and peer review-based evaluations. Both methods are criticised for good reasons, even though the methods applied are usually more sophisticated than described in the caricatures above.

Indicator-driven procedures are usually defined by data availability. The emergence of citation databases has led to the development of many citation-based indicators, often referred to as bibliometrics, that are used to measure research performance. Such indicators are praised for being more objective than expert opinions. However, they have been criticised for being simplistic, neglecting the societal value of research and leading to negative effects on researchers' behaviour (Lane, 2002; MacRoberts & MacRoberts, 2018; Molinié & Bodenhausen, 2010; De Rijcke et al., 2016). The spread of social media led to the development of the so-called altmetrics in the hope that these new indicators could compensate for the disadvantages of bibliometric evaluation, but soon it became clear that similar problems arise: it is not really clear what these indicators measure. There are many different motivations behind citations just as behind downloads, Tweets, reads etc. Citations could refer to criticism, tradition, contrasting or connection to a body of literature (Bornmann & Daniel, 2008; Tahamtan & Bornmann, 2018), while altmetrics could indicate societal uptake, usefulness, satire or decrial, scientific fraud or easy availability of the full text (Bornmann & Haunschild, 2018; Gumpenberger et al., 2016; Lanamäki et al., 2019). In short, the indicators are only loosely linked to the concept they purport to measure: research quality.

The situation is not much better regarding peer review. There are many studies about biases and issues of fairness, inter-rater reliability, predictive validity, subjectivity etc. of peer review (see for example Bornmann, 2011; Langfeldt, 2010; Lee et al., 2013). However, here I want to point to the specific relationship between peer review and the identification of research quality. Usually, peers use in their judgement their own notions of quality that remain implicit. They not only remain implicit to others in the process but are usually also not explicitly known by the expert who judges the work in a holistic way ("I know it when I see it"). Such holistic decisions are, therefore, proven to be inadequate to judge merit (Thorngate et al., 2009) because experts might use a different set of evaluation criteria for each object or at least might apply a different weighting of the criteria between objects of assessment (Langfeldt, 2001). This opens the door for a whole set of biases, such as preference for some schools of thought or methods, conservativism, gender discrimination and so on.

Obviously, both quantitative and qualitative methods for research evaluation suffer from the same problem: It is not clear what they actually measure or identify. What is research quality? It is clear that there is a problem of validity. Validity means here simply the degree to which a measure (that is an indicator or an evaluation) reflects the concept that it purports to measure (Kelley, 1927, p. 14).

While in STEM disciplines, quantitative evaluations did not cause much objection, the situation is different in the SSH. In this chapter, I will argue that the social sciences and the humanities are reflective sciences, understanding the social, historical and political contextuality of research. Rather than trying to "measure" what good research is, they negotiate notions of quality, theoretical and methodological paradigms as well as missions in complex discursive processes and are conscious about the political dimension – and ephemeral validity – evaluation procedures entail (see also Dahler-Larsen, 2012). What might seem erratic and subjective is seen as inevitable part of the process, in fact, as the very essence of research and evaluation. I will show why it is not an easy task to identify research quality and why it is not something universal. But I will also argue that it is still possible. Using well-established methods from the social sciences, I will suggest an approach how to identify research quality taking into account the context and combining indicators and expert judgement.

The chapter is structured as follows: First, I will argue that research quality is a latent construct and introduce some methodological tools how to measure such latent constructs. I then briefly review the research on quality criteria and present a catalogue of general quality criteria for social science research. Combining the methodological approach introduced in the beginning with the quality

criteria, I will present an approach how to measure or identify research quality. I will address some criticisms often brought forward against a methodologically sound and contextual approach to identifying research quality and conclude by pointing out how to identify research quality in practice.

Measuring Latent Constructs

In the social sciences, many concepts researchers want to study are not manifest but rather abstract notions that are nevertheless relevant in our interactions. Such concepts are called latent constructs and include concepts such as intelligence, socio-economic status, work-life conflict or happiness. To be capable of measuring such abstract ideas, we first have to conceptualize them before it can be operationalized with a set of indicators, which is a theoretical task that cannot be replaced by data or statistical methods (Borsboom et al., 2004). This is all the more important as many social concepts are subject to change according to historical and social contexts. Research quality is no exception to that. In fact, it is entirely context-sensitive (e.g., the research quality to select a Nobel Prize winner is not the same as the research quality to attribute a research grant to young scholars). Therefore, research quality cannot be understood as something that can be defined for all times and measured the same way in all contexts. Thus, the often-heard statement that quality should be avoided because it is too complex to define clearly is entirely wrong. Exactly *because* it is context-sensitive, it needs to be defined *explicitly* for each context. This can be done in the following way. In a first step, the dimensions that make up the concept have to be identified. If the concept is complex, the dimensions have to be further specified by aspects until the unit can potentially be measured by one or more indicators. The second step then consists in identifying indicators measuring the aspects.

There are two different forms of relating a concept to indicators: effect indicator models and causal indicator models (see Bollen & Lennox, 1991). On the one hand, and linked to the classical test theory, indicators can be seen as dependent on the concept. When the concept changes, all indicators measuring it also change in the same direction. The indicators are thus "effect indicators", the concept determines the indicators. For example, the concept could be depression and the indicators are general symptoms of depression. This kind of measurement is close to the measurement of manifest concepts (e.g., three measures at different places at two times measure the change of temperature in a room). Figure 1a shows schematically such an effect indicator model. Effect indicators are supposed to be (strongly) correlated as they all reflect the same concept and internal consistency is important for a precise measurement. Indicators that do not correlate enough with the concept are considered inefficient indicators and are excluded from the model. The Crohnbach's alpha is a simple and often used (but heavily criticized) statistic when evaluating an effect indicator model, and factor analysis, especially confirmatory factor analysis (CFI), is used to model the latent construct. Because the indicators reflect the concept, this measurement model is sometimes also referred to as "reflective model" (Fornell & Bookstein, 1982).

On the other hand, and much less known, indicators can make up the concept. The concept changes as one or more indicators building it change. This is called a "causal indicator model" (Bollen & Lennox, 1991) or "formative model" (Fornell & Bookstein, 1982) because the indicators cause or form the concept (see Figure 1b). In this model, the indicators do not need to correlate. A change in one indicator can even be compensated by a change in another indicator. The latent construct is a weighted combination of its observed indicators. In fact, causal indicators should not correlate too much as this would mean that redundant variables are included, and some aspects are therefore overweighted as more indicators measure one aspect than another. Because the indicators define the latent construct, indicators cannot always be dropped from the model as this could change the meaning of the concept. A famous example for a construct measured as a causal indicator or formative model is the socio-economic status, measured by level of education, income and occupation. Note that two persons' socio-economic status can be equally high even if the first person has a lower education, for example, when the first person has a higher income. The two indicators can compensate each other. Similarly, if income would be excluded from the model, the resulting construct would not measure socio-economic status anymore as this is defined by education, income and occupation.

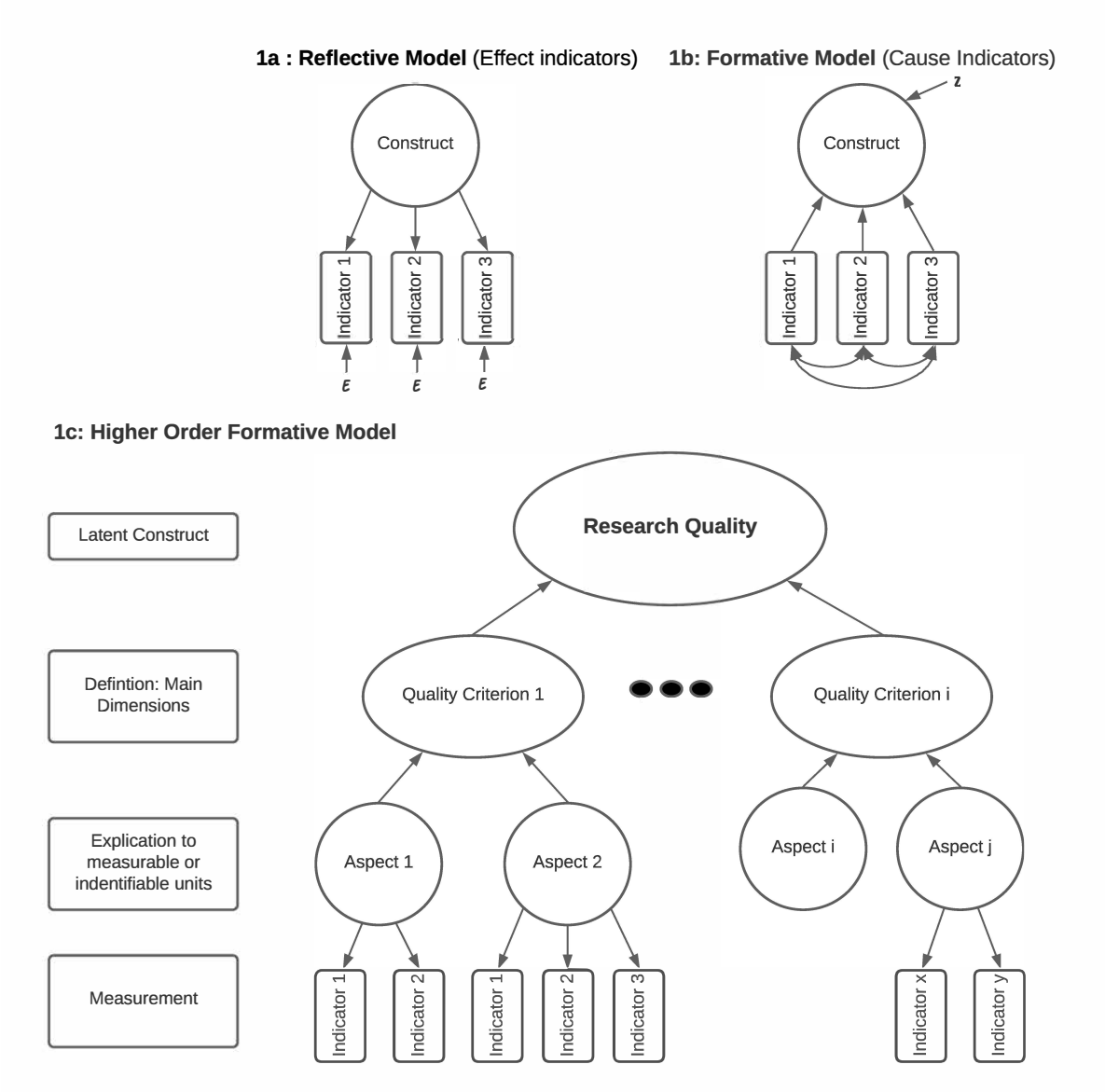

Figure 1. Schematic representation of a conceptualisation of research quality as a higher order formative construct.

The causal indicator model is less used as an explicit measurement model in empirical research as it comes with some issues difficult to resolve: there is a problem of multicollinearity, no item-specific error terms can be defined and it is statistically underidentified, which means that they can only be tested in nomological networks (Diamantopoulos et al., 2008). Therefore, usually, they are used as composite indices. The advantage of composite indices is their capability of summarizing multidimensional constructs and their usefulness for policy analysis (Nardo et al., 2008), their disadvantage is that it is difficult to account for measurement error and that the weighting of the indicators is rarely made consciously, in other words the model is rarely made explicit.

This distinction between the two models is very important. Both models come with very different assumptions and practical implications. However, only a few researchers are aware of this distinction. Often, researchers conceptualize the construct as a causal indicator model, and choose the indicators accordingly but apply an effect indicator model as it is simpler, and the statistical literature focuses strongly on this model (Diamantopoulos et al., 2008; Jarvis et al., 2003). This can lead to quite strong errors in parameter estimation, even if model fit statistics are good: first, because the selection process of indicators is biased, second because direction of causality is inversed (Diamantopoulos et al., 2008).

Certainly, whichever model is chosen, its validity can only be assessed when knowing the aspects of the constructs to measure. For the effect indicator model, we need to know what we measure in order to judge whether the indicators chosen actually reflect the construct sufficiently; for the causal indicator model, we need to have indicators for all relevant aspects in the model as missing one would change the interpretation of the construct.

To define a latent concept, the first step is therefore to identify all aspects that are relevant for the concept. In a second step, indicators can be collected and assigned to the aspects. In the third step, the indicators need to be identified as effect or cause indicators. Figure 1c shows a schematic representation of such a conceptualisation of research quality. On the first level, there are quality criteria that define different dimensions of research quality. These are then specified more clearly by aspects of those criteria. Finally, indicators can be assigned to the aspects. It is possible that for some aspects it is impossible to find indicators, which means that they are only subject to peer evaluation (which can be turned into an indicator by a rating). Note that research quality as a concept is a formative or causal indicator model as it is highly context-sensitive and multidimensional. Criteria form research quality. Criteria themselves are defined by aspects; the aspects, however, can be measured using a reflective or effect indicator model.

One could argue that this is too complicated and that it is likely that you could find two or three indicators that reflect the whole construct of research quality (something also known as g-factor). While it is theoretically possible to find a purely reflective measurement model for research quality, it nevertheless needs to be validated against all dimensions of research quality. This would basically mean that the indicators would correlate strongly with all aspects. Given the highly contextual nature of research quality, it is, however, practically very unlikely to find indicators that reflect the whole construct.

To sum up this methodological section, we can conclude that to measure research quality, we first have to acknowledge that research quality is a latent construct. There are different ways of measuring a latent construct, but in each case, we first need to explicitly know what research quality means in the specific context.

What is research quality?

Thus, the first question to answer is “what is research quality”? We cannot outsource the answer to this question to correlational models or indicators (Borsboom et al., 2004). We can also not revert to the issue that research quality is too complex to define and therefore, only experts can judge it, even if they cannot make it explicit how. If it was too complex, we should accept that we cannot evaluate research. It is also ethically questionable to avoid defining what research quality is by stating that it is too complicated, and thus costly, and research evaluation thus should just be done according to a simple procedure being a “good enough” approximation. We should remind ourselves that research evaluation is judging research and asks the researchers to do high quality research. How, then, can it be ethically justifiable to say that the evaluators can be lazy and apply simple reductionist procedures to evaluate researchers expected to deliver world-leading research?

For these reasons, there have been several attempts in carving out evaluation criteria. However, evaluation procedures usually apply the criteria top-down from a managerial perspective. For example, the UK’s Research Assessment Exercise in 2008 applied the criteria *originality*, *rigour* and *significance*, which are very closely related to Polanyi’s (1962) theoretically developed criteria (originality, plausibility and scientific value). However, a study on how these criteria were applied showed that the interpretations of these criteria by experts across disciplines remained very vague (Johnston, 2008). In Germany (Wissenschaftsrat, 2004) and the Netherlands (VSNU et al., 2003), research *performance* has been defined a bit more fine-grained, yet still remaining a top-down definition with research *quality* being one among other factors and remaining not clearly specified. The definitions provoked strong criticisms from social sciences and humanities scholars as being reductionist and too focused on quantifiable aspects rather than relevant aspects for research, which

led to several adaptations (Hornung et al., 2016; KNAW, 2011). This shows that a top-down definition of research quality (or performance) is difficult because prescribed criteria do not reflect research practices adequately and must remain vague to be adaptable to all kinds of disciplines. When evaluation procedures are at odds with what is considered as good research by the scholars or when important aspects of research quality are missing, negative steering effects are the likely result. If quality is difficult or impossible to define generally, notions of quality are relevant. An alternative would be to take a bottom-up approach and to start from the quality perceptions by the scholars, who can judge research quality best (Hemlin, 1993; LERU, 2012). This is relevant even from a policy point of view: If evaluation aims at improving research – and if this isn't the aim, why bother to evaluate in the first place (Hicks, 2012; Zacharewiczet al., 2018) – capturing the concept of research quality from the researchers' point of view is critical in understanding whether this goal is reached.

There are studies developing quality criteria empirically from how researchers understand or evaluate research (for a short overview, see Ochsner, 2020). From the 1960s to the 1980s, the focus of such studies was placed on what makes a good manuscript. The authors were interested in how editors and reviewers evaluate manuscripts, or how to best teach students how to write scientific articles (Chase, 1970; Frantz, 1968; Wolff, 1970). In the 1990s and 2000s, the studies expanded the scope of quality conceptions from journal articles to evaluation of research grants (Hartmann & Neidhardt, 1990) or evaluation of research in general (Hemlin, 1993) and investigating differences between epistemologically different disciplines (Guetzkow et al., 2004). In the 2010s, the focus shifted to measurement models, criticising that bibliometric approaches reduce research quality to the measurable aspects, proposing more detailed concepts of research quality (Bazeley, 2010; Hug et al., 2013). In recent years, attention has been drawn to how applicants for academic positions are evaluated (Hamann, 2019; Reymert et al., 2020; Van den Brink & Benschop, 2012). This evaluation situation is special in that it entails organisational criteria, such as specific nominations of a subfield or a research tradition. The candidates are also often seen within a specific cohort (Hamann, 2019). However, besides research quality, teaching, administrative tasks, social skills (as colleague and team leader) and language skills play a role (Reymert et al., 2020), informally even physical appearance and age-range (Van den Brink & Benschop, 2012), which makes these studies less relevant for the purpose of this chapter.

What can be learned from those studies is that research quality is multidimensional and context dependent. In a systematic review of such studies inductively developing quality criteria for manuscript and grant review, Hug et al. (2020) find that studies report more criteria for manuscript than for grant review (44 vs. 26 on average). But the number of criteria is high for both. They show that most studies examine medical and health sciences as well as social sciences and that no studies were identified for natural sciences and engineering. Hug and Aeschbach (2020) go a step further and summarise the quality criteria used in the studies on grant peer review identifying 15 evaluation criteria that can be applied to several entities of the object of evaluation leading to quite a complex system of evaluation criteria (for example, the quality criterion *rigour* can be applied to entities such as *theory*, *method* or *conclusions*). Using network analysis, they were able to further systematise the complex system into the broad themes *aims*, *means* and *outcomes* of research. Within the *aims*, the topic and research questions are evaluated according to the criteria *relevance* and *originality*. Regarding the *means*, two aspects are relevant: the research process and the project resources. The first can be evaluated using criteria such as *appropriateness*, *rigour*, or *clarity* with regard to entities such as theory, approach, methods. The second can be evaluated using the *feasibility* regarding resources or the research environment, while *traits*, *motivation* and *diversity* are criteria to evaluate the candidate. Finally, the *outcomes* can be evaluated using *relevance* and *originality* regarding the results (Hug & Aeschbach, 2020, p. 12).

Several studies identify a gap between quality criteria of scholars and those of policy makers and evaluators. Ochsner et al. (2012) show that many quality criteria commonly used in evaluations, such as productivity, recognition, societal impact, relevance, continuity are those least valued by

humanities scholars. Wouters et al. (2014) identify a mismatch between what indicators used in evaluations measure and what researchers' work entails. Wouters (2017, p.109) later defines the "evaluation gap" as the phenomenon that "the criteria in assessments do not match the character or goals of the research under evaluation or the role that the researcher aims to play in society". Langfeldt et al. (2020) differentiate between Field-type (F-type) and Space-type (S-type) notions of quality. The first, F-type notions of quality, originate in the disciplines and are linked to the knowledge production practices in the field. They are "anchored in knowledge pools and/or conditions necessary to enhance these pools" (Langfeldt et al., 2020, p. 119). The second, S-type notions of quality, have their roots in the policy and funding spaces and are held by knowledgeable lay groups, such as policy makers, evaluators, scholars of other disciplines or university administrators. S-type notions of quality are "anchored in considerations exogenous to the field's specific knowledge pools" (Langfeldt et al., 2020, p. 119). The authors also emphasise that, historically, the social and economic contribution of science belong to this type.

Whereas the diagnosis of those studies is similar, they are quite different in nature. Wouters et al. (2014) identify that indicators cannot measure all relevant aspects of research and that interpretations of why an indicator is used might differ, Langfeldt et al. (2020) claim that there are different ideas of what is relevant, Ochsner et al. (2012), finally, point to the fact that indicators and criteria used most often in evaluations are not differentiating between better or less good research; or, in the terminology of Langfeldt et al. (2020), that if S-type criteria are implemented without reference to F-type criteria, they will most likely lead to "unintended" effects.

Research quality in the social sciences

As presented in the methodological section, to measure a latent construct such as research quality, we need a conceptual measurement model, consisting of quality criteria specified by potentially measurable aspects (see Figure 1c). Starting from the comprehensive study on general notions of research quality in the humanities that fulfils this requirement (Ochsner et al., 2014), Ochsner and Dokmanović (2017) refined the criteria catalogue to match social science research. The team consisting of scholars from sociology, psychology and law found that most criteria identified for the humanities might also apply to the social sciences. However, some adaptations had to be made. Some changes in wording were necessary, e.g., the aspect "stringent, comprehensible and convincing" was changed to "stringent, reproducible and convincing" to reflect the more quantitative nature of social sciences research. Some new aspects were introduced, for example "be bold to raise issues which are vital for the society" to reflect the potentially problematic relationship with politics and society, especially in countries with restricted scientific freedom. The criterion *impact and relation to society* was divided into *impact on society* and *relation to society*. This change drew from the result found in the humanities that there is a clear distinction between the two (Ochsner et al., 2013), but it was deemed even more important for the social sciences as social sciences research relates directly to the current society. Instead, the criterion *fostering cultural memory* was dropped to avoid confusion with the two criteria *relation to* or *impact on society* and because it seemed that fostering cultural memory is rather a humanities-specific criterion.

All in all, a criteria catalogue of 19 criteria further specified by 65 aspects resulted from the discussion (for the criteria see Table 1, for a full list of all aspects, see Table A1). The list of criteria was tested inviting all social scientists in Macedonia to rate the criteria (the response rate was 26%). We found that most criteria were relevant to Macedonian social scientists with the exception of *scholarly exchange* and *impact on society*.

The results again confirmed that there is a clear difference between *relation to society* and *impact on society* (Ochsner et al., 2013). The *relation to society* is considered important, i.e., addressing topics that the researchers perceive as important for society, vulgarising research results to communicate it to the public, and interacting with the profession linked to the discipline. However, *impact on society* is not considered a quality criterion for research, i.e., to address topics that the public perceives as

important and to have an actual effect on the society. This might have to do with the fact that it is very difficult to ascribe impact to a specific research as pathways to impact are diverse (Muhonen et al., 2020) and research might even have negative impact if there is a too strong focus on achieving and proving such an impact (Derrick et al., 2018).

Table 1. Criteria for research quality in the social sciences

1. Scholarly exchange*	8. Impact on research community	15. Scholarship, erudition
2. Innovation, originality	9. Relation to society	16. Passion, enthusiasm
3. Productivity	10. Impact on society*	17. Vision of future research
4. Rigour	11.Variety of research	18. Connection between research and teaching, scholarship of teaching
5. Recognition	12. Connection to other research	
6. Reflection, criticism	13. Openness towards ideas and persons	19. Relevance
7. Continuity, continuation	14. Self-management, independence	

Note. Criteria consist of 65 aspects. If all aspects of a criterion did not reach consensus, i.e. more than ten per cent of the scholars disagreed that the aspect can be used to adequately assess research, the criteria is considered not having reached consensus and is marked with a star (*).

This section has shown that research quality is indeed multidimensional and that different evaluation situations ask for different quality criteria. If research quality is to be measured, it has first to be defined according to the evaluation situation. Only in the next step, indicators can be assigned to the relevant quality criteria and aspects.

Measuring research quality

Having a clear idea on what research quality means in a specific context, i.e. selecting the relevant quality criteria and aspects (see Figure 1c), we can attempt to measure it. Ochsner et al. (2014) followed such an approach for the construct "institutional research evaluation of experienced researchers in the humanities". They identified 19 quality criteria specified by 70 aspects (Hug et al., 2013). In an encompassing literature research covering scientific and grey literature as well as evaluation protocols and a survey, Ochsner et al. (2012) identified such a large number of indicators potentially applicable to this construct that they aggregated them into 62 indicator groups. They assigned them to the aspects and criteria and found that 50% of the aspects that have reached a broad consensus among the humanities scholars cannot be measured quantitatively. More importantly, those indicators that are commonly used in evaluation procedures, such as citations, third-party funding, prizes, transfer to economy or society, measure exactly those criteria that were judged least important by the scholars, with the exception of two criteria, showing the "evaluation gap" mentioned above. Ochsner and Dokmanović (2017) identified this gap also for the social sciences, however to a somewhat less pronounced degree. It is thus no surprise that researchers identified a number of negative steering effects of the application of quantitative research assessments (De Rijcke et al., 2016).

Whereas it is most important to start from the scholars' notions of quality in order to reduce the risk for negative steering effects, very often other stakeholders' notions of quality also play a role. For example, in an institutional evaluation, the institution's mission becomes relevant; or for a certain funder, the effect regarding its statutory purpose is important. In fact, with the rising importance attributed to societal impact, the term "stakeholders" is often used but has mainly been reduced to actors in economy and the public (Spaapen et al., 2007). However, given the early meaning of

stakeholder as reconstructed by Freeman and Reed (1983, p. 91), even the narrow definition includes "any identifiable group or individual on which the organization is dependent for its continued survival". Thus, a limited understanding of stakeholder as (some vaguely defined) external groups is problematic because it covers the interests at play of several groups involved in evaluation, such as the evaluators, data providers, science policy makers and, importantly, the scholars. Ochsner et al. (2020) identify four types of stakeholders: research production; research consumption and use; research policy and administration; and evaluation services. Each of the four groups contains many interest groups. While this does not mean that all possible stakeholders' notions of quality must be included in each evaluation situation, it helps identifying the relevant stakeholders and also possibly identifying where some hidden interests come into play, such as commercial interests of data providers.

Applying a rigid modelling approach and starting from the scholars' notions of quality does not mean that the notions of quality of other stakeholders are ignored. Quite the opposite. They can be made explicit and be integrated into the catalogue of criteria. The advantage of such an approach is that it becomes explicit what are the notions of quality of the different stakeholders, thus attributing value to different points of views while still staying true to the actual purpose of evaluation: identifying quality of research.

For example, while science policy makers prefer criteria like internationality, interdisciplinarity and impact on society, these are not seen as quality criteria for research by scholars as it exists bad international, interdisciplinary research just as it exists good national and disciplinary research (Ochsner et al., 2013 for the humanities and Table A1 in the appendix for the social sciences). These are thus not really *quality* criteria but rather *policy* priorities, valorising some types of research more than others. Disguising them as quality criteria might lead to communication problems and unacceptance among scholars. Nevertheless, they surely can be a legitimate part of an evaluation procedure. However, they have to be applied *on top* of quality criteria, i.e., first, the quality needs to be assured and then, those policy criteria might further select research.

To sum up, measuring research quality in a context-dependent way is possible and feasible. Being explicit about criteria applied in an evaluation, differentiated by stakeholders, enhances not only the validity of the measure but also its transparency and the communication during the evaluation process.

Critical discussions of the approach

Research quality is a concept regarding a social issue. It is logical, then, to apply social science methodology to capture it. Even more so, as previous methods have proven to be inadequate or problematic (De Rijcke et al., 2016). Why is this approach so rarely applied? There are several arguments brought forward by evaluators and bibliometricians to counter such a model-based approach. First, it is considered too complicated; second, bibliometric approaches might not work on the individual level but they work on the aggregate level; third, there are two sorts of conflicting quality criteria, F-type (researchers) and S-type (policy makers); fourth, context-specific concepts lack comparability; fifth, there is a difference between identifying quality (i.e. criteria-based evaluation) versus excellence (i.e., indicator-based, comparative evaluation); sixth, epistemic traditions in a discipline guide evaluation and this approach applies only to few disciplines, like the humanities. In the following, I will address these criticisms (or rather defences of the status quo) and show why the complexity is an advantage rather than a disadvantage.

First, the argument that such an approach is not practical as it is too complicated, counters ethical expectations to the profession of evaluators. If evaluation is judging which research is of high quality and which is not, probably having financial or even personal consequences, it is difficult to say that evaluators themselves do not need to apply a rigorous procedure. If all research is only legitimised if it is of outstanding quality, its evaluation needs to be precise enough as well to meet such a goal: to identify high quality. Ethical aspects of research evaluation are still under-researched and certainly not considered enough in practice (Biagetti et al., 2020).

Second, bibliometricians often argue that research metrics should not be applied on the individual level but that they work well on the aggregate level (Mittermaier, 2021; Russel & Rousseau, 2009). However, a proof why an invalid measure on the individual level would become valid by aggregation is still missing. A simple example shows that aggregation will not make a flawed indicator valid. If in an election, only half of the votes are counted for female candidates while for men, all votes are counted, the vote counts are not valid on the individual level. But if we aggregate from the individual to the party level, or from community to country level, the election results will still be biased, women will get less candidates elected and parties with more female candidates will get less seats. In other words, and with an example for evaluation, if social science research is not well represented in a data base like Web of Science (Van Leeuwen, 2013), aggregation of citation counts per field to the country level will still not validly represent how social science research relates to natural science research across countries. The aggregated citation count for social science is simply biased as the source is already biased.

Third, it is often argued that there are different quality perceptions between scholars and policy makers and their notions of quality are conflicting. Therefore, they must be separated into so-called F-type (field) and S-type (space) quality criteria (Langfeldt et al., 2020). While it is important to acknowledge the different interests and notions of quality, the juxtaposition comes with the risk of claiming that S-type criteria are important in evaluations as it is about policy or allocating scarce resources, not about what is considered interesting in the field. This is a dangerous misconception. F-type and S-type notions of quality are not independent of each other. To achieve S-type quality, F-type quality needs to be assured, the goal of any evaluation is finally to improve research; low-F-type-quality research will not lead to high S-type quality. Therefore, the model suggested here does not juxtapose the two but rather emphasises that there are different ideas of what good research is. It is all about the communication between different stakeholders (and not just two!), but at the core, there is always F-type research quality as it is the fundament of any quality concept; S-type notions of quality are legitimate, yet not alternatives but complementary to F-type criteria. S-type notions of quality need to be translated, so that they become internalised by researchers. When S-type notions of quality cannot be translated into F-type criteria, they are political or societal and are thus not quality criteria but other aspects of research that are relevant, *on top* of the quality of research. Their evaluation is not a question of quality but a political negotiation.

Fourth, one critique is that if research quality is defined in a context-specific way, results of evaluations cannot be compared. However, what is the use in comparing across evaluations? Any evaluation has its own goal. We can compare different units according to the goal "world-leading basic research in nursing". Compared regarding another goal, say "provide the region's hospitals with applied knowledge and trained specialists", the result of a comparison of the same two units might be very different. Both comparisons, if made explicitly for these goals might be perfectly valid even if the outcome is the opposite. However, comparing the institutions according to a third goal not relevant to either question, say "visibility in rankings", would make the result invalid regarding the first two questions. Being context-specific does not mean that comparison becomes impossible, rather, it gives meaning to the comparison. Dahler-Larsen's (2012) seminal work provides detailed explanations of this issue.

Fifth, it is often argued that there is a difference in judging quality comparatively, putting forward the notion of excellence, versus judging quality criteria-based, focusing on notions of quality. It is argued that, on the one hand, it might be that the selection processes at universities work well, and only good research is produced. Still, some research is better, outstanding or excellent. For such purposes, we do not need to be aware of all criteria but are interested in comparing quantitatively the performance according to bibliometric and scientometric indicators. On the other hand, if we want to judge whether a research article can be published or a degree awarded or not, quality criteria need to be applied to identify whether the research fulfils scientific standards. While this difference is indeed relevant because they represent two very different evaluation situations, the question arises what then

excellence means if the indicators used to identify excellence are not validated, i.e. are not representing research quality? If we think that soccer is a team sport and that a good player needs to be able to read the game, have team spirit, communicate well on the field, handle the ball well etc., we could argue that in the highest league, there are only good players, so to identify the most excellent players, we cannot use complicated measures but take the time a player is in ball contact, as dribbling is an important skill and being in ball contact means being in control. Would that measure help us to identify the most excellent players? And what would be the consequences for the team sport if the best players would be selected according to this indicator? Obviously, to identify excellence among the good, we still need to know explicitly what is "better" and validate our measure against this concept.

Finally, there is the argument that different disciplines follow different epistemic patterns, i.e. how they produce valid knowledge. Therefore, their evaluation should differ according to these epistemic foundations of the disciplines (Bonaccorsi, 2018). Indeed, this is crucial regarding the fact how one can measure certain aspects of research quality, because the epistemic traditions guide publication and dissemination practices and, therefore, indicators need to reflect such differences. For example, if in philosophy complex issues are described using long, encompassing arguments in books, and relevant findings are published in such books, an evaluation based only on citations from journal articles will miss the most important sources of citations and hence not produce valid results. However, one cannot go as far as turn the argument around: natural sciences "measure" their objects and therefore use data-driven quantitative measurements to produce knowledge, consequently, research evaluation needs to be data-driven and quantitative to be valid (Bonaccorsi in this volume). Similarly, economics and psychology adopted the methods from the natural sciences and use mainly empirical results and communicate more in articles than in books, and therefore, the argument is that their research can be evaluated as is done in the natural sciences. If this was so, in fact, one should follow the argument to the end and claim that if research evaluation needs to use the logic of the object to evaluate it, then, this should be true for the research methods applied in the disciplines as well: regarding natural sciences, the object is nature, nature uses random trial and error processes in a survival of the fittest model. Therefore, natural sciences should not use logic and mathematics to study nature but should generate erratically random laws and look which laws survive the longest. Or economics studies economic principles, thus economic research should follow economic goals: the research result receiving the highest price on the market will be the truest. Obviously, this is nonsensical as the method should not reflect the methods used by the object of inquiry but should be the best methods to answer the research question in the field. Related to this chapter's topic on research quality in the social sciences, we cannot argue that just because economics mimic natural science methods, we can use the same evaluation procedures as are used for the natural sciences, while anthropology not mimicking natural sciences needs another approach. Economics are rarely describing ahistorical law-like phenomena and are just as locally oriented as anthropology (or even more so); it is crucial for economics to include those aspects into the evaluation procedure just as in anthropology. Furthermore, there is no prove that bibliometrics is a valid tool for evaluating research in the natural sciences. Rather, the evaluation of research is a social phenomenon whether natural sciences or social sciences are evaluated and has its own epistemic principles. Therefore, the methods applied in research evaluation need to be able to capture social phenomena; they must, of course, respect the particularities of the fields, such as different epistemic traditions.

These points show that the complexity and context-specificity is a strong advantage of the approach presented above. It makes visible the complexity of research and makes it possible to address the diversity of situations research evaluation has to take into account.

Conclusion: Identifying research quality in the social sciences in practice

The evaluation of research is a practice related to social sciences. Identifying research quality as a major goal of research evaluation therefore needs to follow social sciences methodology. Research quality is a complex, latent construct, i.e. it cannot be directly observed. Furthermore, it is subject to

context: There is no absolute research quality, but the quality can differ between regions, across time and evaluation situations. We have seen that both current practices, using either a quantitative approach, such as bibliometrics or scientometrics, or peer review come with validity issues as they conceptualise the construct as a reflective model when a formative model should be applied.

In this chapter, I suggest thinking of research quality as a higher-order formative construct, that is composed of its relevant dimensions (see Figure 1c). Following the terminology of Hug et al. (2013), research quality is defined by a set of quality criteria (dimensions) that can be further specified by aspects (measurable units), for which then indicators can be identified. Some of the aspects will not be open to quantitative measurement but can only be judged by experts.

The aim of any research evaluation is to improve research quality or to select the best from less worthwhile research (Hicks, 2012). It is therefore crucial to start the conceptualisation of research quality with the scholars' notions of quality as the evaluation will feed back on their research activities. There are several studies that identify scholars' notions of research quality in the social sciences for several evaluation situations (see for overviews Hug et al., 2020; Hug & Aeschbach, 2020). However, there are many stakeholders involved in research (Langfeldt et al., 2020). Therefore, the notions of quality of the relevant stakeholders need to be taken into account (Ochsner et al., 2020). Besides notions of quality, stakeholders might bring in other concepts relevant to the endeavour of research, such as internationality, interdisciplinarity or societal relevance. These concepts are orthogonal to research quality (local research can be just as good as international research can be bad) and should be evaluated separately[2].

After having defined the concept by explicating quality criteria and aspects, indicators can be identified that can inform about the aspects of research quality. For some aspects, no indicators will be found.

Finally, experts can judge each aspect of research quality separately, taking into account information of indicators if available. It is important, that experts rate each aspect separately. Thus, it can be assured that for each object of evaluation, the same weighting of the different aspects is used (Thorngate et al., 2009).

While such an approach seems to be complicated, it has a number of advantages: It is centred around the scholars' notions of quality and thus focuses on what is the ultimate goal of evaluation: judging the quality of research. However, it adds notions of quality of other stakeholders, makes them explicit and puts them in relation to the scholars' notions of quality. This enhances communication, reduces the risk of boycott or negative steering effects and makes the political dimension of evaluation explicit. The last point is crucial as evaluation is "inherently political. It aspires to help democracy work" (Dahler-Larsen, 2012, p. 234) through social sense making. Making the notions of quality in an evaluation exercise explicit is thus a legitimizing and democratic act.

Acknowledgements

[2] Note that some aspects of internationality and interdisciplinarity do form aspects of research quality. For example, research using interdisciplinary sources is usually considered more complete than research not taking them into account – while still being disciplinary research as such. The same holds for international sources. These are aspects of scholarly exchange or connection to existing research. What is not part of research quality is interdisciplinary or international research in terms of the nature of the research (or the topic): research is not of high quality just because the research team consists of persons from several countries or disciplines or because the topic is on an international level rather than a local one.

This chapter draws on the work from the author's work with Hans-Dieter Daniel and Sven E. Hug and from the COST Action CA 15137 'European Network for Research Evaluation in the SSH (ENRESSH)'. The author would like to express gratitude for all discussions on research quality with the members of these projects but remains solely responsible for the content of this chapter. The adaptation of the quality criteria to the social sciences and the survey presented in this chapter were conducted by Mišo Dokmanović and Michael Ochsner during a Short Term Scientific Mission of the COST Action CA 15137 'European Network for Research Evaluation in the SSH (ENRESSH)'. Sven E. Hug contributed to the discussions on the adaptation.

References

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education, 35*(8), 889-903. http://doi.org/10.1080/03075070903348404w

Biagetti, M. T., Gedutis, A., & Ma, L. (2020). Ethical Theories in Research Evaluation: An Exploratory Approach. *Scholarly Assessment Reports, 2*(1), Article 11. http://doi.org/10.29024/sar.19

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314. http://doi.org/10.1037/0033-2909.110.2.305

Bonaccorsi, A. (2018). Towards an Epistemic Approach to Evaluation in SSH. In A. Bonaccorsi (Ed.), *The evaluation of research in social sciences and humanities: Lessons from the Italian experience* (pp. 1-29). Springer. http://doi.org/10.1007/978-3-319-68554-0_1

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1), 197-245. http://doi.org/10.1002/aris.2011.1440450112

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Bornmann, L., & Haunschild, R. (2018). Allegation of scientific misconduct increases Twitter attention. *Scientometrics, 115*(2), 1097-1100. https://doi.org/10.1007/s11192-018-2698-6

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071. http://doi.org/10.1037/0033-295X.111.4.1061

Chase, J. M. (1970). Normative Criteria for Scientific Publication. *The American Sociologist, 5*(1), 262-265. https://www.jstor.org/stable/27701631?seq=1

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford University Press.

De Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use: A literature review. *Research Evaluation, 25*(2), 161-169. http://doi.org/10.1093/reseval/rvv038

Derrick, G., Faria, R., Benneworth, P., Budtz-Petersen, D., & Sivertsen, G. (2018). Towards characterizing negative impact: Introducing Grimpact. In R. Costas, T. P. Franssen, & A. Yegros-Yegros (Eds.), *STI 2018 Conference Proceedings* (pp. 1199-1213). CWTS.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203-1218. http://doi.org/10.1016/j.jbusres.2008.01.009

Fornell, C., & Bookstein, F. L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research, 19*(4), 440-452. http://doi.org/10.1177/002224378201900406

Frantz, T. T. (1968). Criteria for publishable manuscripts. *Personnel and Guidance Journal, 47*(4), 384-386.

Freeman, R. E., & Reed, D. L. (1983). Stockholders and stakeholders: A new perspective on corporate governance. *California Management Review, 25*(3), 88–106.

Gozlan, C. (2016). Les sciences humaines et sociales face aux standards d'évaluation de la qualité académique. *Sociologie, 7*(3), 261-280. http://doi.org/10.3917/socio.073.0261

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is Originality in the Humanities and the Social Sciences? *American Sociological Review, 69*(2), 190-212. http://doi.org/10.1177/000312240406900203

Gumpenberger, C., Glänzel, W., & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics, 108*(2), 977-982. http://doi.org/10.1007/s11192-016-1991-5

Hamann, J. (2019). The making of professors: Assessment and recognition in academic recruitment. *Social Studies of Science, 49*(6), 919-941. http://doi.org/10.1177/0306312719880017

Hartmann, I., & Neidhardt, F. (1990). Peer review at the Deutsche Forschungsgemeinschaft. *Scientometrics, 19*(5-6), 419-425.

Hemlin, S. (1993). Scientific quality in the eyes of the scientist: A questionnaire study. *Scientometrics, 27*(1), 3-18.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251-261. http://doi.org/10.1016/j.respol.2011.09.007

Hornung, A., Khlavna, V., & Korte, B. (2016). Research Rating Anglistik/Amerikanistik of the German Council of Science and Humanities. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research assessment in the humanities. Towards criteria and procedures* (pp. 219–233). Springer. http://doi.org/10.1007/978-3-319-29016-4_17

Hug, S. E., & Aeschbach, M. (2020). Criteria for assessing grant applications: A systematic review. *Palgrave Communications, 6*, Article 37. http://doi.org/10.1057/s41599-020-0412-9

Hug, S. E., Hołowiecki, M., Ma, L., Aeschbach, M., & Ochsner, M. (2020). Practices of peer review in the SSH I: A systematic review of peer review criteria. In M. Ochsner, N. Kancewicz-Hoffman, M. Hołowiecki, & J. Holm (Eds.), *Overview of peer review practices in the SSH* (pp. 61-66). ENRESSH. https://dx.doi.org/10.6084/m9.figshare.12032589

Hug, S. E., Ochsner, M., & Daniel, H. D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369-383. http://doi.org/10.1093/reseval/rvt008

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research, 30*(2), 199-218. http://doi.org/10.1086/376806

Johnston, R. (2008). On Structuring Subjective Judgements: Originality, Significance and Rigour in RAE2008. *Higher Education Quarterly, 62*(1-2), 120-147. http://doi.org/10.1111/j.1468-2273.2008.00378.x

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book.

KNAW. (2011). *Quality Indicators for Research in the Humanities*. Royal Netherlands Academy of Arts and Sciences.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Harvard University Press.

Lanamäki, A., Ahmad, M. U., & Ochsner, M. (2019). *Any publicity good publicity? The effect of satirical bias on Twitter and the Altmetrics Attention Score* [Conference presentation]. Book of abstracts of the 3rd RESSH Conference, Valencia, Spain. https://ressh2019.webs.upv.es/wp-content/uploads/2019/10/ressh_2019_paper_40.pdf

Lane, J. (2010). Let's make science metrics more scientific. *Nature, 464*(7288), 488-489. http://doi.org/10.1038/464488a

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science, 31*(6), 820-841.

Langfeldt, L. (2010). Expert panels evaluating research: Decision-making and sources of bias. *Research Evaluation, 13*(1), 51-62.

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. *Minerva, 58*, 115-137. http://doi.org/10.1007/s11024-019-09385-2

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17. http://doi.org/10.1002/asi.22784

MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology, 69*(3), 474-482. http://doi.org/10.1002/asi.23970

Mittermaier, B. (2021). Peer review and bibliometrics. In R. Ball (Ed.), *Handbook bibliometrics* (pp. 73-86). De Gruyter.

Molinié, A., & Bodenhausen, G. (2010). Bibliometrics as Weapons of Mass Citation. *CHIMIA International Journal for Chemistry, 64*(1), 78-89. http://doi.org/10.2533/chimia.2010.78

Muhonen, R., Benneworth, P., & Olmos-Peñuela, J. (2020). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. *Research Evaluation, 29*(1), 34-47. http://doi.org/10.1093/reseval/rvz003

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on constructing composite indicators: Methodology and user guide*. OECD.

Ochsner, M. (2020). Messung von Forschungsleistungen? Was gemessen wird und was gemessen werden will. In I. M. Welpe, J. Stumpf-Wollersheim, N. Folger, & M. Prenzel (Eds.), *Leistungsbewertung in wissenschaftlichen Institutionen und Universitäten* (pp. 350-378). De Gruyter. http://doi.org/10.1515/9783110689884-017

Ochsner, M., & Dokmanović, M. (2017). Quality criteria and research obstacles in the SSH in Macedonia. In *Book of abstracts of the Second international conference on research evaluation in the social sciences and humanities* (pp. 69-71). University of Antwerp.

Ochsner, M., Hug, S. E., & Daniel, H. D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(2), 79-92. http://doi.org/10.1093/reseval/rvs039

Ochsner, M., Hug, S. E., & Daniel, H. D. (2014). Setting the stage for the assessment of research quality in the humanities: Consolidating the results of four empirical studies. *Zeitschrift Für Erziehungswissenschaft, 17*(6), 111-132. http://doi.org/10.1007/s11618-014-0576-4

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for Research Quality in the Humanities: Opportunities and Limitations. *Bibliometrie - Praxis und Forschung, 1*, Article 4.

Ochsner, M., Ma, L., Kancewicz-Hoffman, N., Holm, J., Gedutis, A., Šima, K., Hug, S. E., Dewaele, A., & De Jong, S. (2020). *Better Adapted Procedures for Research Evaluation in the SSH*. ENRESSH. http://doi.org/10.6084/m9.figshare.12049314

LERU. (2012). *Research Universities and Research Assessment*. Position Paper for the League of European Research Universities. https://www.leru.org/files/Research-Universities-and-Research-Assessment-Full-paper.pdf

Polanyi, M. (1962). The Republic of Science: Its Political and Economic Theory. *Minerva, 1*(1), 54-73.

Reymert, I., Jungblut, J., & Borlaug, N. S. B. (2020). Are evaluative cultures national or global? A cross-national study on evaluative cultures in academic recruitment processes in Europe. *Higher Education*. Advance online publication. http://doi.org/10.1007/s10734-020-00659-3

Russell, J. M., & Rousseau, R. (2009). Bibliomterics and institutional evaluation. In R. Arvanitis (Ed.), *Science and Technology Policy* (Vol. 2, pp. 42-64). Oxford.

Spaapen, J., Dijstelbloem, H., & Wamelink, F. (2007). *Evaluating research in context: A method for comprehensive assessment* (2nd ed.). Consultative Committee of Sector Councils for Research and Development. http://www.eric-project.nl/files.nsf/pages/NWOA_73VH8D/$file/eric_book_internet.pdf

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*(1), 203-216. http://doi.org/10.1016/j.joi.2018.01.002

Thorngate, W., Dawes, R. M., & Foddy, M. (2009). *Judging merit.* Psychology Press.

Van den Brink, M., & Benschop, Y. (2012). Gender practices in the construction of academic excellence: Sheep with five legs. *Organization, 19*(4), 507-524. http://doi.org/10.1177/1350508411414293

Van Leeuwen, T. N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: A problematic relationship? *Bibliometrie - Praxis und Forschung, 2*, Article 8.

VSNU, NWO, & KNAW. (2003). *Standard Evaluation Protocol (2003-2009) for Public Research Organisations*. https://www.knaw.nl/en/news/publications/standard-evaluation-protocol-2003-2009-for-public-research-organisations/@@download/pdf_file/90000091.pdf

Wissenschaftsrat. (2004). *Recommendations for rankings in the system of higher education and research*. http://www.wissenschaftsrat.de/download/archiv/6285-04-engl.pdf

Wolff, W. M. (1970). A study of criteria for journal manuscripts. *American Psychologist, 25*(7), 636-639. http://doi.org/10.1037/h0029770

Wouters, P. F. (2017). Bridging the Evaluation Gap. *Engaging Science, Technology, and Society, 3*, 108-118. http://doi.org/10.17351/ests2017.115

Wouters, P. F., Bar-Ilan, J., Thelwall, M., Aguillo, I. F., Must, Ü., Havemann, F., Kretschmer, H., & Schneider, J. W. (2014). *Acumen: Final report*. Acumen. https://vdocuments.net/download/finalreport29april2014-29042014-acumenfinalreport29april2014

Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2018). Performance-based research funding in EU Member States: A comparative assessment. *Science and Public Policy, 46*(1), 105-115. http://doi.org/10.1093/scipol/scy041

Appendix

Table A1. Criteria and aspects for research quality in the social sciences, their ratings and consensus across scholars in Macedonia

Criterion	Aspect	Mean	Median	P10	Consensus
Scholarly exchange	Disciplinary Exchange	4.86	5	3	0
	Interdisciplinary Exchange	4.37	5	3	0
	International Exchange	4.70	5	3	0
Innovation, originality	Innovation regarding data	5.10	5	4	1
	Introduction of new research topics	5.18	5	4	1
	New approach to topic or data	4.95	5	4	1
	Generating new paradigms	4.94	5	4	1
	Contribution of new findings or interpretations	5.05	5	4	1
	Innovation regarding methodology	4.76	5	3	0
	Identification of gaps in existing knowledge	5.16	5	4	1
Productivity	Continuous research outputs	4.89	5	4	1
Rigour	Systematic and transparent research process	5.21	5	4	1
	Stringent and reproducible argumentation	5.10	5	4	1
	Presentation of relevant documents and evidence	5.50	6	5	1
	Clear language	5.52	6	5	1
	Clear structure	5.55	6	5	1
	Reflection of method	5.28	5	4	1
	Adhere to rules of scientific honesty	5.71	6	5	1
	Discussion of generalizability of insights	5.26	5	4	1
Recognition	Insights are recognized by the research community	4.82	5	4	1
	Insights are recognized by society	4.71	5	4	1
	Reputation within research community	4.84	5	4	1
	Reputation in society	4.68	5	3	0
	Reputation at own university	4.98	5	4	1
	Question the notion of 'definitive and final truth'	4.76	5	4	1

Reflection, criticism	Criticizing assertive claims and social norms fashionable in the current society	4.74	5	4	1
	Criticizing established scholarly approaches	4.30	4	3	0
	Self-critical and self-reflective research	4.68	5	3	0
	Disclose complexity in society	4.87	5	4	1
Continuity, continuation	Promotion of young academics	5.08	5	4	1
	Continuation of research traditions	4.77	5	3	0
	Long-term pursuit of research topics	4.08	4	3	0
Impact on research community	Stimulating new research	5.17	5	4	1
	Establishing a new school of thought	4.33	4	3	0
	Research used in work of others	5.02	5	4	1
	Strategic decisions impact the research community	4.54	5	2	0
Relation to society	Topics relevant for society from the scholars' perspective	5.12	5	4	1
	Conveying findings to a non-academic audience	4.97	5	4	1
	Impact on profession related to discipline	5.12	5	4	1
	Bold to raise questions relevant to society but might have negative consequences on career	4.81	5	3	0
Impact on society	Responding to societal concerns	4.72	5	3	0
	Effect on national, regional or local culture/society	4.61	5	3	0
Variety of research	Contributing towards variety and diversity	4.97	5	4	1
	Taking risks and working outside of mainstream	4.69	5	3	0
Connection to other research	Building on current state of research	4.93	5	4	1
	Re-connecting to neglected research	4.17	4	2	0
	Engaging in ongoing research debates	4.82	5	4	1
Oppenness towards ideas and persons	Openness to other, competing ideas	5.31	5	5	1
	Openness to other persons	5.40	5.5	5	1
	Realization of own research goals	5.38	5	5	1

Self-management, independence	Open-ended, unpredictable research	5.09	5	4	1
	Research is not directly utilizable	3.69	4	2	0
	Research is not directly targeted at a recipient	3.75	4	2	0
Scholarship, erudition	Rich experience with data, sources, materials	5.30	5	4	1
	Knowledge based on own research	5.35	5	4	1
Passion, enthusiasm	Passionate about research or the topic	5.25	5	4	1
	Arouse passion for research or the topic	5.16	5	4	1
	Intrinsic motivation for research activity	5.01	5	4	1
Vision of future research	Pointing out important research for the future	5.09	5	4	1
Connection between research and teaching, scholarship of teaching	Research-based teaching	4.56	5	4	1
	Own research affected by teaching	4.29	4	3	0
	Research has its impact mainly in teaching	3.93	4	2	0
	Building character of oneself and others	5.00	5	4	1
	Social competency	5.27	5	4	1
Relevance	Research is relevant for the research community	5.14	5	4	1

Note. Aspects are assigned to criteria and listed in the order of presentation in the questionnaire. Reported statistics are the mean, median and the 10^{th} percentile of all participants. Consensus is reached (reported as "1") if the aspect is rated at least with a median of 5 and a 10^{th} percentile of 4.