Norms, Quality and Dissonance

Theory and Practice of Science According to Scientists

Corinna Elsenbroich

Centre for Research in Social Simulation, University of Surrey

Maria Xenitidou
Centre for Research in Social Simulation, University of Surrey

Alastair Gill

Digital Humanities, Kings College London

Abstract

How do we determine what is good quality science? Or how should we determine what is good quality science? Quality in science can be characterized from different vantage points and always has descriptive as well as normative aspects. We present data on conceptions, understandings and negotiations of quality in science. Stipulating three models of quality in science and embedding them into institutional practices, we analyse different notions of quality. We conclude that different models create tension and dissonance for practicing scientists who negotiate personal notions of quality individually rather than institutionally.

Keywords: sociology of science, scientific discourse, normativity

1. Introduction

In 2010 the *Lancet*, a high profile medical journal, withdrew a paper published in 1998 that linked the MMR[1] vaccine to autism and bowl disease. In the 12 years since the original paper, many studies were conducted but none replicated the findings. The paper appears to report falsehoods rather than truth. This on its own, however, is not a reason for the withdrawal of a paper from the scientific literature. In fact falsifying scientific findings is the stuff science is made of (e.g. Popper (1934/1959)). There were several problems with the study such as the use of ethically not approved procedures and a biased selection of participating children. In 2010 the Lancet withdrew the paper from publication.

> 'Following the judgment of the UK General Medical Council's Fitness to Practice Panel on Jan 28, 2010, it has become clear that several elements of the 1998 paper by Wakefield et al are incorrect, contrary to the findings of an earlier investigation. In particular, the claims in the original paper that children were "consecutively referred" and that investigations were "approved" by the local ethics committee have been proven to be false. Therefore we fully retract this paper from the published record.'[2] (The Editors of the *Lancet*)

Wakefiel was struck off the professional register by the General Medical Council after being found guilty of more than 30 charges (Triggle 2010). These charges included

- Failure to disclose how patients were recruited for the study.
- Conducting the stud without approval from the hospital's ethics committee
- Purchasing blood samples from children present at his son's birthday party.

---

[1] Measles, Mumps, and Rubella.

[2] The earlier investigation referred to is Hodgson (2004).

- The research being financed by a team of lawyers representing parents suing vaccine manufacturers in the US.

So 'false' science can remain in the public sphere as long as it is 'good quality' false science, i.e. it conforms to scientific quality and ethics standards. In this article we argue that there are different 'models' of scientific quality of which we will discuss three.

The first model we want to discuss is the *metric model*. In this model quality is captured in (mainly quantitative) measures, applied in for example in university league tables[3], academic recruitment, the allocation of funding, etc. Although many countries use some measure of scientific output, for the metric model we focus on the United Kingdom as it is currently the most measured academic system (Gray and Jenkins 2004). Other European countries are starting to follow in the UK's footsteps, making it ever more important to assess the metric model and its potential impact on quality in science.

The second model we discuss could be called the *ideal model*. In this model, science is defined as a set of norms, roughly the norms identified in Merton (1973). Rather than judging the outcomes of a scientific endeavour what matters is that scientific findings are arrived at in the right and proper way. We refer to the third model as the *practice model* as it is extrapolated from what scientists say and do with regards to quality in their daily (research) life. While we present the three models as distinct, we expect a certain level of interaction between them. The metric model takes into account the ideal and practice models, the practice model is definitely 'aware' of the metric model and the ideal model was developed with a sense of practice science. What this interaction consists of we discuss later in the paper.

---

[3] In the UK different organisations publish university league tables which are based on the outcomes of the past Research Assessment Exercise (RAE) and the upcoming Research Excellence Framework (REF) plus a range of other measures, e.g. the National Student Survey, spending per student, job prospects of graduates.

We will compare these models, point to overlaps, discrepancies and tensions and look at the scientific and administrative processes that support or undermine each model.

The analysis of these different models of quality in science is of interest given the explosion in scientific literature due to a proliferation of journals and web based publishing (Gill and Gilbert 2009), increased competition for funding accelerated by financial cuts, the involvement of private sector funding and increased administration of universities. Amidst the churn of these developments, it seems that different actors in science – researchers, administrators – find themselves in different places, with different roles and goals (see Burrows 2012).

In order to understand these roles better within the contexts they emerge, the paper is structured as follows: in Sections 2 and 3 we present theoretical analyses of the metric model and the ideal model. In section 4, we summarize data on how scientists see quality in their field, practice and community and develop the practice model of science from this data. In section 5, we evaluate how scientists' perceptions and understandings of quality align with the discussed models and identify tensions within scientists' understandings of quality.

2. Measured Science

What is quality? Quality is interdefinable with the notion of 'standard'. Something is `of quality' if it reaches a certain standard and a standard is an agreed level of quality. In turn, inherent in the notion of standard is the idea of measurement. Hence, in order to assess the quality of something, a standard and its measurement needs to be defined. For a car a standard could be defined from a multi-dimensional chart including a car's price, speed, fuel efficiency etc. as a set of proxies for quality. Each one of the attributes can be measured independently and compared between cars. An overall standard might be defined using a weighting of the different, independently measured proxies. Things become more complicated if an attribute like 'beautiful' is added to judge the quality of a car. Although beauty might be of great importance, it arguably lies in the eye of the beholder and is difficult or impossible to measure.

If we want to decide on quality in science we also need to think about the standard, the proxies and the relevant measurement. 'Truth' could be seen as a useful proxy for quality in science. Given the purpose of science to describe and explain the world, truth seems to be an essential ingredient of good science. Although an important proxy, similar to the beauty of a car, we cannot 'measure' the truth of a piece of research. Many scientific findings persisted as truths for a long time before being falsified and superseded (see Kuhn 1968). The attempt to find formal criteria that are truth indicative failed with the failure of logical positivism, or as Ayer put it in an interview in the 1970s: 'I suppose the most important [defect of logical positivism] . . . was that nearly all of it was false' (Hanfling 1981).

Given the problem with truth one might want to take *impact* as a proxy for quality in science. There are two kinds of impact on which science is judged, academic and non-academic impact. We will discussed the former below in Section 2.1 and focus here on non-academic impact.[4] Although real world impact is certainly a success of science, measuring this kind of impact is riddled with problems. Scientific advances often impact much later than and in completely different ways than intended or expected at their conception (e.g. Laser, Internet, Teflon). Secondly, it is not clear what might count as impact. One might mean engineering output such as microwave ovens and non-stick pans or a wide intellectual influence such as the theory of psychoanalysis. It might be something that is used by many, such as the Internet or rocket ships, which are used by very few but culminate in a moon landing. Thirdly, we might need a lot of 'low impact science' to

---

[4] See Smeyers and Burbules (2011) for a discussion of the potential tension between 'real world impact' and academic impact.

produce high impact result, or as Newton put it, a lot of science is 'standing on the shoulders of giants'. Given these problems, impact as a proxy does thus not fare much better than truth.

Decisions about research are however made, such as which articles or books to publish, how to allocate research funding and who to hire or promote. One input into these decisions is usually expert knowledge, in particular in the form of peer-review for articles, books and research proposals. Using expert knowledge as the basis for quality assessment could be seen as following the principle of 'governance'. It is the purpose of the governance of the scientific profession to ensure its output is of high quality. However, governance has fallen from grace in the last decades and has largely been replaced by administration for many professions, most notably education and health (Goldstein and Spiegelhalter 1996). Similarly, science has undergone an administrative turn since the 1980s. Once governance is replaced by administration, it is no longer experts judging quality but administrators without expert knowledge. This means measures for quality have to be found that are easily accessible and comparable.

Current administrative quality assessments are based mainly on publication and citation measures, leading to a quantitative measure of scientific output.

2.1 Publication and Citation Measures

For a long time academia has been a 'publish or perish' industry (Garfield 1986) but it seems that the measurement of quality of scientific output using publication and citation indexes has made the academic world an even harsher place (Fanelli 2010). Citation can be seen as a proxy for the impact of a paper and, in turn, a researcher or a journal. There are two major publication metrics we want to discuss here, the *h-index* relating to researchers and the *Impact Factor* relating to journals.

The h-index, *h,* is intended to measure both the productivity and impact of a researcher or research group. *h* is the number of papers published by an author with *h* or more citations (Hirsch 2006). Hirsch asserts that it is the best single measure of a person's scientific achievement and should be used to compare researchers on the same level of seniority. For comparing researchers at different stages, the index is adapted to reflect the career stage by dividing the h-index by the number *m,* the number of years since the researcher's first publication. The h-index has some positive features, such as taking into account the potentially long tail distribution of citations. It is, however, flawed as a simple measure to compare scientists. Too much information is lost in reducing a scientist's achievement to the number *h*. In Adler et al. (2008) intuitive examples are given to show that the h-index is far from useful for the comparison of researchers' achievements (p. 11). One problem is that all publications are discarded that have citations below *h*, meaning a scientist with 10 papers, each with 10 or more citations has the same *h* as a scientist with 100 papers, 10 of which have 10 or more citations, the remaining 90 having 9 citations. Another problem is that the number of citations above *h* is also ignored, meaning that a scientist with 10 papers, one with 10 citations and 9 with 100 citations, has the same *h* as a scientist with 10 papers all of which have 10 citations. Thirdly, the value of the h-index is heavily dependent on the database used. For example, Thomson-Reuters' *Web of Knowledge* and Elsevier's *Scopus* give lower scores than *Google Scholar* because the former only covers selected journals (Bar-Ilan 2008). *Google Scholar* in turn has the problem that on the web, author's names are not unique identifiers, leading to false attributions of publications and citations. These shortcomings are particularly problematic if the h-index is used as the primary justification for important scientific decisions, such as recruitment, funding, and career advancement (Burrows 2012; Adler et al. 2008).

There are also more general questions about whether citation is an appropriate measure of the quality of a publication. To think so makes several assumptions about the nature of citation as a positive appreciation of the content of a paper. But there are different motives for citing. A paper might simply be cited for a mistake, in disagreement or even warning, i.e. a *negative* citation. Again, just counting citations is not sensitive to this use. Different kinds of papers get different numbers of citations. Very general and review articles get more citations than articles reporting new or very specific research findings without the former being of any better quality than the latter, just aimed at a broader set of readers. Thus, citations are neither an indication of having thoroughly examined (or even read) the cited paper, nor that the outcome of this examination was positive (c.f. 'negative' or 'warning' citations).

Whilst the h-index is supposed to measure the quality of a researcher or research group, the *Impact Factor* (IF) is the equivalent measure for the quality of journals. It is calculated as the mean number of citations to articles in a journal over the previous two years.

Citation distributions are extremely skewed making the arithmetic mean a not very useful measure. The following example of an article entitled "A short history of SHELX" shows the sensitivity of the IF to outliers. The article was published in 2008 and received 2391 citations in 2009. In the case of the SHELX paper mentioned above, the IF of *Acta Cristallographica A* rose from 2.051 in 2008 to 49.926 in 2009. The paper contains the following sentence, searchable in the abstract, providing at least a partial explanation for the high number of citations:

> "This paper could serve as a general literature citation when one or more of the open-source SHELX programs (and the Bruker AXS version SHELXTL) are employed in the course of a crystal-structure determination." (Sheldrick 2008 p.112)

But does this mean that those articles published in *Acta Cristallographica A* in 2009 are of 25 times higher quality than those in 2008? Still, a higher mean number of citations can be seen as a measure of the impact of the journal, and that was what the Impact Factor (IF) was introduced to measure (Garfield 1972). However nowadays the IF is used to judge the quality of single publications or researchers and as that, the IF is not a good measure at all and Garfield, warned against such a use of it in his initial article.

To clarify the problem of judging single publications by the IF of the journal they are published in, Adler et al. 2008 turn the question from averages to probabilities. When comparing articles from two different journals what is the probability that a randomly selected paper from the higher IF journal has more citations than a paper from the lower IF journal? They compare the *Proceedings of the American Mathematical Society* and the *Transactions of the American Mathematical Society*. Using the 2005 database of citations, the *Proceedings* had an impact factor of 0.434, the *Transactions* an impact factor of 0.846. However, drawing a random article from the *Transactions*, 62% of the time it will not have more citations than a randomly drawn article from the *Proceedings*, (see Adler et al. 2008 p.11).

The allure of the IF is easy to see: Given that the IF is a proxy for the impact of a journal why should it not also serve as a proxy for the impact of the *content* of the journal? The above discussion, however, shows how fallacious this inference turns out to be.

2.3 Measurement and Quality

Whereas governance, for example in the form of peer review, needs experts to judge, the main idea of the metric model of quality in science is to construct metrics that can be applied to scientific output by non-experts[5].

---

[5] According to Matzlich (1982) for example, the measuring of science served exactly this purpose: providing indicators of quality to compete for state funding. This dates back to the post 1960s competition for resources occurring as a result of the frame of nation-state administration becoming the ‘source’ of these resources. Criteria for evaluation (science indicators) for resource distribution were based on economic indicators, and were thus mainly quantitative. This became a driver of the professionalization of science-expertise, as classification became a prerequisite of measurement (see Matzlich 1982). Thus, the metric model succumbs the biases of classification rather than the biases of understanding.

We have seen citation measures and impact factors as examples. The metric model is driven by competition; researchers need to increase their h-index *in comparison to other researchers,* getting a higher position in university league tables, such as *Times Higher Education* World University Rankings (intrinsically comparative) or publishing in journals with higher IF *than other journals.*

We have discussed several measures of quality in science that are currently used to make decisions such as what gets published, where to publish and who is recruited. The measures consist of simple mathematical applications of the arithmetic mean or the simple counting of publications with a certain number of citations. Although this simplicity make them perfect for administrative purposes, we have discussed the pitfalls of such simple measures. Assessing quality in science seems to be a more complex endeavour than these metrics allow to measure.

3. Ideal Science

In this section we introduce an alternative, idealised view of science, following Merton's norms of science. Merton's influential work on the interactions of science and culture produced a set of norms of science (Merton 1973). These norms present a set of criteria that demarcate science from other social endeavours. Research violating any one of these norms should not be deemed science (Gilbert & Mulkay 1984; Pickering 1992). Merton identified the following norms as uniquely specifying science: Communality, Universalism, Disinterestedness and Organised Scepticism. Although Merton conceived of the set of norms simply as demarcating science from other social processes, these norms can also assure quality of scientific output as they are all geared towards objectivity. Universalism and Disinterestedness eliminate bias in the research process. For example the problem of bias resulting from self-interest, created from competitive academia is discussed in Fanelli (2009 2010). Communality and Organised

Scepticism support objectivity directly and are closely linked to a consensus theory of truth (Habermas 2003).

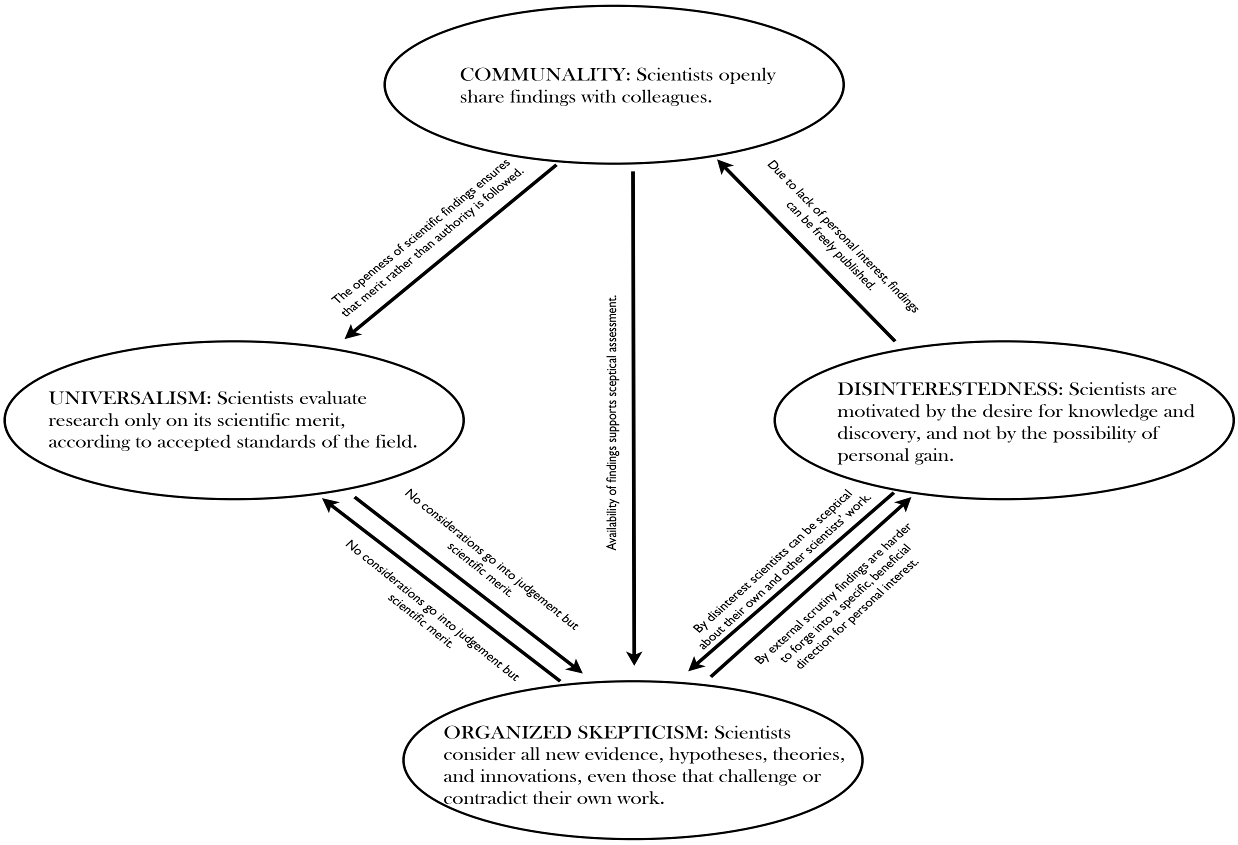

Figure 3.1: The support network between Mertonian norms of science.

These norms mutually support each other and are also supported by processes within the science community.

For example, most scientific work is peer reviewed, i.e. assessed by other scientists to test whether it is worthy to be taken into the body of knowledge. This process supports organised scepticism as well as the communality of scientific research. The peer review process is usually ‘doubly blind’: both the reviewers and the authors remain anonymous. The anonymity of the authors supports the norm of universality because the research will be accepted on its scientific merit, not on the reputation or personal characteristics of the authors. The anonymity of the reviewers supports the norm of disinterestedness, e.g. it reduces the possibility that the reviewers' verdict will be based on favouritism. Recent research has confirmed the importance of disinterestedness for the peer review process. Squazzoni et al. (2012) ask whether peer review should be remunerated. They perform experiments mimicking the processes of peer review and find that any external reward for peer review would be detrimental to the quality of the review. This finding is mostly explained by the 'crowding out effect' of external rewards (cf. Frey 1994), because it makes peer review no longer a norm-driven but a gain-driven activity.

Peer review is used in many scientific processes such as publication and funding decisions. Research funding is particularly sensitive to the norms of disinterestedness. Whereas public funding is usually seen as 'safe', more and more research is funded by private companies leading to a potential conflict of interest, e.g. suppression of results contrary to the company's interest. But public funding can also pose problems to researchers. For example, Fanelli (2009) reports on the apparently high readiness to falsify research findings, in particular among young researchers (post-docs). In the biomedical sciences at the University of California, San Diego, 81% of post-docs reported a willingness 'to select, omit or fabricate data to win a grant or publish a paper' (Fanelli 2009 p. 9).

Other processes supporting Mertonian norms are open access to publications, data, code etc., freedom of movement of academics, and a unified scientific language.

4. The Practice Model

In the previous two sections, we introduced two models of quality in science. The metric model constitutes an attempt to measure scientific output for the purposes of resource-allocation (economic level) and the ideal model refers to how science should and should not be (deontic level). The metric model could be seen as a sub-model of ideal science in the sense that it lays out the general principles (output - 'originality, significance and rigour'; impact - 'reach and significance'; environment – structures supporting research, e.g. collaborations, staff development, research students, income, infrastructure and facilities), processes (research outputs, expert review), standards (e.g. up to four research outputs) and mechanisms (environment, including those assessed and those entitled to assess, e.g. RAE itself) for assessing science. If this were the case, the metric model would need to support the ideal model. Yet, rather than a supportive relationship, there is a direct tension between the models, for example between the disinterestedness norm and the metric performance demands. It has been shown that measurement and proxies generally lead to target hunting rather than quality production thus changing the essence of what they set out to measure (Smeyers and Burbules 2011 p. 10).

Along the same lines Burrows (2012) discusses the succession of assessment exercises and argues that the early ones, up to 1996 were merely trying to provide some transparency for funding allocation but that sometime between the 1996 and 2001 exercise the culture changed from being an assessment to feeding back into the research culture. As Scott (2012) argues, the first assessment led to weeding out 'those self-regarding academics whose great monograph always seemed to be just out of touch at the end of the rainbow' but since then led to game playing rather than research improvement.

Besides the tension between the ideal and metric models, Anderson et al (2007) identified some dissonance between scientists' endorsement of normative standards in science (the ideal norms) and their own behaviour. In a survey of 3247 researchers some normative dissonance was found between scientist's endorsement of normative standards in science and their own behaviour and substantial normative dissonance was found between perceptions of scientists' own behaviour and that of other scientists.

The authors take the list of Mertonian norms of science and add two of their own, elicited as relevant from previous focus group research. They juxtapose the set of six norms with the corresponding counternorms (see table 3.1). The two added norms are particularly interesting for our study. The first norm is Governance together with the counternorm of Administration. This can roughly be described as the processes by which the smooth running of science is ensured. Governance covers processes such as peer-review, open access, self-regulation. Administration, in contrast, is a top down process. Administrators, not necessarily *au fait* with science, govern the processes via targets and regulation. The second norm is Quality and the counternorm is Quantity. These two norm-counternorm pairs are directly concerned with quality in science, governance and administration as different processes to guarantee quality and quality and quantity as juxtaposed outputs of the scientific enterprise. The two are also related as often governance will lead to quality (for example peer review does not consider how many papers should be accepted, but only whether a paper is good enough to be accepted), whereas administration will lead to quantity, given that administration more often than not relies on the measurement of outputs. Substantive dissonance is found in particular regarding the need for high quantity output, leading to other norms, such as communality and disinterestedness, being undermined. Norms in science, just like all other norms, are ideals rather than descriptions of actual behaviour. However, complete non-compliance with norms leads to a diminishing of their strength (Fischerbacher and Gächter 2010). In the case of norms of science non-adherence also has another consequence: If science is defined via a set of norms, widespread non-adherence to the norms leads to the outcome no longer resembling 'science' in the same way as changing the rules of how pieces move on a chess board results in one no longer playing chess. While this is not problematic *prima facie* as first, the interpretation of the 'game' is dynamic and second, the new game might make better sense to what the players' goals are, it may become problematic if it leads to chaos (for example owing to not being based on 'shared' understandings or not being a game at all anymore). In this sense, norm adherence is crucial for the continuation of the 'game' but also crucially depends on norms matching shared understandings about how the game should be played. It is also sensitive to a sense of commitment in order for the game to be played – the perception of others adhering to norms. If I think that most people litter I will feel less inclined not to. It is thus an interesting finding that there is some dissonance between the attitude and behaviour of researchers with regards to ideal norms but also that researchers see themselves as mostly compliant to norms in science and other researchers as mainly non-compliant (Anderson et al 2007).

COMMUNALITY: Scientists openly share findings with colleagues.	SECRECY: Scientists protect their newest findings to ensure priority in publishing, patenting, or applications.
UNIVERSALISM: Scientists evaluate research only on its merit, i.e., according to accepted standards of the field.	PARTICULARISM: Scientists assess new knowledge and its applications based on the reputation and past productivity of the individual or research group.
DISINTERESTEDNESS: Scientists are motivated by the desire for knowledge and discovery, and not by the possibility of personal gain.	Self-INTERESTEDNESS: Scientists compete with others in the same field for funding and recognition of their achievements.
ORGANIZED SKEPTICISM: Scientists consider all new evidence, hypotheses, theories, and innovations, even those that challenge or contradict their own work.	ORGANIZED DOGMATISM: Scientists invest their careers in promoting their own most important findings, theories, or innovations.
GOVERNANCE: Scientists are responsible for the direction and control of science through governance, self-regulation and peer review.	ADMINISTRATION: Scientists rely on administrators to direct the scientific enterprise through management decisions.
QUALITY: Scientists judge each others' contributions to science primarily on the basis of quality.	QUANTITY: Scientists assess each others work primarily on the basis of numbers of publications and grants.

Table 3.1 Norms and Anti-norms influencing scientists' work.

Anderson et al (2007) identified some normative dissonance between scientists' subscription to ideal norms and their own (reported) behaviour. While the tensions noted above might be seen as problematic for the good practice of science on many levels (i.e. indicating confusion about what the good practice of science is leaving it open to interpretation, as well as lack of communalism or shared understanding of what the good practice of science is – both with various repercussions), we argue that that their existence does not necessarily mean the practice of *bad* science, but indicates that there is a need for creating the 'space' for the expression of a model of science that makes sense to all user communities involved.

In the following section we present what we call a practice model of science, elicited from survey and interview data with scientists about quality in science and consider the ways in which it corresponds or 'interacts' with Mertonian norms and science metrics as discussed in Section 2.

5. Scientists' understandings of quality in science

In the Sections 2 and 3 we discussed the ways in which science is currently measured as well as what is considered as 'ideal science'. We then considered empirical data which identified a dissonance between the ways science is perceived and the ways it is practiced. We explore this further in the next sections by drawing on two sets of data on scientists' perceptions and understandings of quality: a medium scale survey, and 20 individual interviews. These were followed up by a focus group study, which we consider briefly at the end.

The research was initiated by an exploratory stage which consisted of running a word elicitation study with the aim to explore how people describe quality and to compare this to previous findings relating to quality perception in other contexts (Ghylin et al. 2008). This was then followed by a survey collecting words associated to quality and then interviews with scientists focusing on their understandings of quality in science.

5.1 The Survey

In addition to the data collection technique used by Ghylin et al. (2008), we also draw upon approaches from psychology (e.g., Rosch 1978; Kearns and Fincham 2004, who used 'prototype' studies to explore concepts such as 'forgiveness').

Words relating to quality were collected using an online survey, run between August and October 2010. The questionnaire asked participants to 'write down the words and phrases that you associate with quality in science' (participants were also asked to perform the same task considering quality in a general context, but we do not discuss these finding in the current paper).

Participants were encouraged to produce both positive and negative examples, and the online questionnaire enabled participants to produce an unlimited number of responses. On the following page of the online survey, participants were then presented with the quality words/phrases that they had previously supplied, and were then asked to rate each on a 7-point scale ranging from 3 ‘Very Positive’ to -3 ‘Very Negative’. The questionnaire also asked participants to complete a short demographic questionnaire.

Participants were recruited via UK universities through a personal network of contacts of the authors, with participants encouraged to distribute information to friends/contacts inside and outside universities. Students, staff, and members of the public were all encouraged to participate, in order to get a broad range of views on quality, especially in science (i.e. by including participant who are ‘inside’ and ‘outside’ science; cf. Matzlich 1982). Of the 225 participants who took part in this study, 170 provided age information (86 aged 21-30 years, 51 less than 21 years, 18 were 31-40 years, 12 were 41-50 years, 2 were 51-60 years, and 1 was 61 years and over); Of the 169 providing gender information, 90 were female, and 79 male; Of the 168 participants providing information about their highest level of education, for 74 this was a school-leaving qualification, for 43 this was an undergraduate degree, for 27 a taught post-graduate degree, and for 27 a research postgraduate degree.

5.1.1 Data processing

The 226 participants gave a total of 955 responses (words/phrases), a mean of 4.23 responses each. Of these responses, 750 were unique, with 90 shared by two or more participants (the most frequently response, ‘reliable’, was provided by 23 individuals). 361 of responses were single (or hyphenated) words, whereas 389 were phrases, and in total the average number of words per response was 2.53.

As noted above, each of the responses was evaluated by the same participant on the following page of the survey on a scale of 3 ‘Very Positive’ to -3 ‘Very

Negative'. The most popular rating was 3 (Very Positive), which was given to 657 responses, followed by 0 (Neutral, 193 responses). The most popular negative rating was -3 (Very Negative), applied to 46 responses, followed by -2 (30 responses), and -1 (29 responses).

Top positive items	ave	freq
Reliable	3	23
Accurate	3	21
Valid	3	7
Useful	3	6
Precise	3	6
Unbiased	3	5
peer reviewed	3	4
Insightful	3	4
Safe	3	4
Relevant	3	4
Logical	3	4
Beneficial	3	3
Verifiable	3	3
Excellence	3	3
Practical	3	3
well written	3	3
Reliability	3	3
Replicable	3	3
Consistent	3	3
quality assurance	3	3

Top negative items	ave	freq
Confused	-2.5	2
Faulty	-2.5	2
Plagiarism	-2.5	2
Bad	-2	7
Sloppy	-2	2
Unverified	-2	2
Poor	-1.7	3
Inconclusive	-1	2

Table 4.1 Showing the most strongly rated positive and negative items which relate to quality in science

5.1.2 Coding of data

The next step of the processing of the data was to group items according to their semantic similarity to help to give insights on the data analysed in the following stages of this study. The Wmatrix corpus comparison tool was then used to identify (Rayson 2008) set phrases and multi-word expressions, with texts parsed using n-gram software (Banerjee and Pedersen 2003) to identify less common quality-related 2-, 3-, and 4-grams. N-grams which did not relate to content (e.g., ‘of the’) and duplicate items were removed by hand. Items with a frequency of at least 2 were retained, giving a total of 312 science-related quality items. Grouping of the items into categories was performed by two researchers with expertise in text and content analysis, with both having an understanding of the literature relating to science and quality in order to take into account the context of the items under consideration (see Strauss and Corbin 1998). The researchers coded the data separately, with differences in categories resolved through mutual agreement (Charmaz 1994; Lofland and Lofland 1995). This process resulted in 23 science quality topic categories.

Science Quality Topic Categories	Description	Number of items
appearance	appearance, elegant, presentation	6
clarity	well communicated, understandable, clear arguments	14
context	context, subject, topic	3
correctness	accurate, precise, exact	19
depth	insight, thorough, detailed	17
ethics	unbiased, ethical, safe	11
evaluation/assessment	good, excellence, high	16
function/result	usable, practical, beneficial	14
information-oriented	information, informative,	2
intelligence	intelligent, smart, clever	3
new	ground breaking, significant, surprising	22
new/relational	different	1
process-oriented	well designed, analysis, conditions	42
proof-oriented	believable, convincing	2
quantity	mass, quantity	2
relational	cited, contribution, recognised	13
resilience	durable, stands the test of time, solid	5
results/proof	testing, verifiable, accurate data	50
standards	conforms, guidelines, requirements	10
structure-oriented	arguments, designed, framework	16
trust-oriented	reputation, rigour, confidence	18
trust-oriented/category	professional, academic, expertise	6
value	value, worth, expensive	11

Table 4.2 The Scientific quality word groupings

5.1.3 Discussion of Survey

As can be seen from the results for both the most positive and negative items of the survey, and also from the human coded categories, what constitutes quality in science is largely framed using Mertonian norms. For example, 'reliable' and 'accurate' are by far the most popular positive descriptions of quality science; of the top 20 words presented in table 4.1, only `peer reviewed', 'practical', and 'quality assurance' can be seen to hint at any kind of external assessment of scientific quality. In the coded categories, we similarly find categories overwhelmingly reflecting a Mertonian perspective (correctness, depth, evaluation/assessment), but to a lesser extent also find items hinting again at external assessment (function/result, new, standards, and value).

5.2 The Interviews

Twenty 30-minute semi-structured interviews with scientists from Physical & Engineering Sciences in the UK were conducted between October and December 2010. Participants included professors, senior lecturers, lecturers, research fellows and research students. The interviews focused on quality as a collective process and on the mechanisms and structures through which quality in science is produced and constituted (Boltanski and Thevenot 2006; Latour 2007). They were recorded and transcribed, each producing approximately 3,500 words of text. Transcriptions were subsequently discourse analysed focusing on regularities and irregularities in themes and arguments as well as the rhetorical strategies used and footing employed by participants (Wetherell 1998).

5.2.1 Discussion

As we will see by focusing on the interview data, the practice model of science is supported by some processes in place to measure science (see extracts 1-3 below), supported by new processes (see extracts 4a & b, 5 below) and not supported (anti-supported) by some existing processes (see extracts 3-

5 below). The extracts discussed below are presented as exemplar cases of the above. Participants range from senior lecturers to professors and also the majority also holds administrative roles (including one Head of Department, two Deans and one Associate Dean and Director of Research Centre).

As regards the first indication coming from the data, there seems to be some alignment with the 'ideal' sense of quality (vs quantity) as an evaluative standard. For example, participants would argue that publishing extensively is not an indication of good quality (but rather the opposite, see extract 1 below) and would be judgmental of colleagues or research cultures that practice this (see extract 2).

It is noteworthy that extract 1 comes from an interview with a participant who also has an administrative role in the faculty.

Extract 1 (Interview 13)

1 R: [..] in in physical sciences to publish two or three papers a year is totally
2 normal
3 I: mm
4 R: e:h but to publish sort of three or four good papers in a five year period
5 that's what marks you out as a as a good researcher I think. ((and)) It's
6 difficult to judge which ones are the good ones unless you are right in that
7 field but I'm that's what I think should be what people should care more
8 about is °the quality rather than the volume°

The participant orients to science as divided – 'in physical science' – and to publishing two or three papers a year as 'normal' (line 1). Yet, he distinguishes between papers and 'good' papers, which he treats as a criterion of one's own merit in research. The criterion or standard is: three or four in a five year period, which is in accordance with the standard set in the metric model (see Section 2.2). While the standard corresponds to the metric model in terms of the number of output, the quality – 'good' – requires field expert judgment (line 5). In the deontic that follows quality as 'that' is vaguely constructed as ‘three or four good papers in a five year period’ and is ‘that’ which is juxtaposed with quantity (‘volume’).

Extract 2 (Interview 11)

R: I referee quite a lot of papers that from groups is go: >not trying to be a racist or xenophobic or anything but< there’s a lot of groups in China who: publish a lot of papers on very similar things like: there’s a lot of repetition a:nd I think they have a: very strong drivers there to publish as much as they can a:nd I receive a lot of papers that I look at and say “well this has been done fifty times before” I could find identical fifty papers
I: mm
R: showing exactly the same thing
I: mm
R: a:nd keep on they keep on sending them [...] there’s no use no kind of driver to: making something making something better out of this it’s just that “we get that and we then go and do something else”. The kind of churn of publications rather than actually being on a project and try to go towards
I: mm mhm
R: steps of advancement. I’m sure I’m as equally as guilty for a lot of this as many other people w((h))ere you are driven by the number of publications you put out
I: mm mhm
R: I think higher quality is often when you don’t just keep the churn of (.) material coming focusing on making something better o:r improving what you’ve done

The extract seems to support the findings in Anderson et al (2007) with regards to normative dissonance within and between scientists. According to this, not only subscription to norms is somewhat higher than norm adherence (lines 15-6), but other scientists' behaviour is considered as much more counter-normative than one's own. However, in lines 15-6, the speaker might be seen as disclaiming or warding off potential accusations of prejudice as was the case in lines 1-2, thus managing accountability rather than admitting a mismatch between attitude and behaviour with regards to quality.

In the next extracts, the ways in which the quality (vs quantity) norm has been applied indicate an uneasy relationship between scientists and metrics (extracts 3, 4a&b, 5 below). For example, in extract 3 a tension between established measures of quality assessment and measurement, and 'true' quality is indicated, challenging the foundations of 'universalism'.

It is noteworthy that extract 3 comes from an interview with a participant who mainly has an administrative role in the faculty.

Extract 3 (Interview 7)

66 R: I mean the good example recently that won the Nobel prize in physics (.)
67 these two guys in Manchester e:h discovered (.) graphene. A:nd fine >you
68 know< the implications for that are very profound. This is wonderful
69 material it could revolutionise electronics a:nd best conductor known to
70 mankind it's just e:h extraordinarily strong material etcetera etcetera ok?
71 That's fine. They did it by a bit of cello tape and pencilled it (.) ok (.) is that?
72 anyone can do that. °an eleven year old kid can do that°. but that's not really
73 what it's about. (...) So is that high quality? (......) e:h the work is published in
74 the high quality journals (.) the: spin offs are enormous (.) absolutely huge.
75 It's one of the big areas of research (.) in a few years
76 I: mm
77 R: ok? So there's countless examples of this (.) in in in science in general and
78 °in physics in particular°. A:h so yeah that's what I associate with quality.

79 Otherwise it’s impossible to judge what is what good quality is. I mean you
80 do read a paper and you think “yeah this is superbly written”
81 I: mm
82 R: “it’s very clear, it’s thorough e:h it’s these people have done a good job”.
83 You can tell. They’ve sort of covered all the bases e:h and you say this is a
84 high quality paper. It’s well written (.) well-presented. E:h (..) but I think in
85 general I’d consider quality to be a little bit more than that. °it’s what it is (.)
86 the relevance of the work°

Extract 3 displays a construction of quality in science through a subtle contrast on which (and only on which) the speaker grounds (the) resort to official recognition of the scientific work as a quality assessment metric.

The exchange unfolds in response to Question 2 asking the interviewee to provide an example of high quality in their field.

The exchange is torn between quality constructed as scientific work having profound implications (impact,) on the one hand, and well-conducted, well-written and well-presented scientific work, on the other, and, on this distinction the speaker grounds his association of quality based on established, ‘recognisable’ measures of quality (Nobel Prize, published in high quality journals, impact) disclaiming this grounding on that ‘otherwise it is impossible to judge what good quality is’.

Namely, while in lines 66-69 the speaker constructs quality as scientific work having profound implications, in lines 70-72 the speaker shifts footing and focuses to the scientific work entailed ‘behind’ producing science with profound implications. This formulation eventually escalates to a rhetorical question: ‘so is that high quality?’ (line 72). In response to the question the speaker draws on a list of ‘accepted standards’: published in high quality journals (output), enormous spin offs (impact), one of the big areas of research (impact), all of which subscribe to the metric model.

The formulation oscillates between the norm of 'universalism' – evaluating research on its merit through accepted standards, the accepted standards corresponding to the metric model and to particularism – (questioning) the processes behind the output – and orients to the former as 'it is impossible to judge quality otherwise'. To this end the speaker offers an alternative example of quality in science, which is constructed as the opposite, emphasizing the research process (rather than the 'measurable' impact). Yet, this version of quality is not missing the 'measurable' aspects *per se* but rather, a third criterion: 'the relevance of the work' (line 84).

Therefore, the speaker oscillates between the norms of universalism (applied through processes and principles laid out in the metric model) and the norm of particularism (emphasizing the research process 'behind' the output), and offers a third option as a solution in practice (see also extract 5). In this way, the speaker also manages accountability by presenting a balanced argument and by disclaiming the questioning of 'accepted scientific standards'.

In extract 4a the participant orients to quality by drawing an explicit distinction between 'external' and 'internal' measures. The account seems to align with norms of ideal science with regards to 'quality' (vs quantity) but challenges the 'accepted scientific standards' which are 'in place' to represent the norm, thus the metric sub-model.

Extract 4a (Interview 11)

I: and how do you develop cause you said you said there's the personal a:nd how do you develop this sort of your own understanding? How do you learn about what quality is

R: ((laughs))

I: and develop your own understanding? What's this (..) process about?

R: a:m (..) >I don't know I think< for me personally one of the biggest drivers was actually one of my old bosses who: when I was a post doc at the university of Bristol. Where the in that group the focus was on doing something well rather than pushing out publications °my boss there was not very very interested in° publishing in particularly high impact factor journals o:r particularly pushing out as many papers as possible. He really just liked doing things correctly and you'll all rather than saying "well that looks fine (.) out it goes" he always said "could we do this extra or could we do that extra" so it was more: focused towards bei:ng sort of accurate. Am (....) conversely sort of here now funding ((laughs)) being much tighter these days and a lot more focus is put on the: kind of meeting the metrics that are being set by the funding bodies and the journals so the aim is maybe the driver there maybe when funding is tighter you more sort of judge yourself by the external metrics when i:t's (.) when you are left on your own you maybe more going towards your own internal sort of quality drivers ((inaudible)) what you want to do

Internal quality drivers seem to correspond to ideal norms, disinterestedness and organized skepticism in particular, while the external metrics, as set by the funding bodies and journals (part of the metric model), imply self-interestedness and dogmatism.

Extract 4b (Interview 11)

The extract then continues with the speaker explicitly arguing that ideal science is best represented by internal metrics that each scientist has than by 'accepted scientific standards' (the metric model) which are 'in place' to represent science.

I: mm. But from what you are saying they are not necessarily incompatible a:s
R: no::
I: at odds wi:th
R: no: (.) not necessarily you sort of what you want to do does overlap as to what you submit
I: mm

26 R: but i:f if say the only way you could see if you wanted to do a study of
27 something and you need to do it pretty carefully and do a thousand (..) run
28 thousand samples and do it this way and it will be quite formulaic maybe
29 that (..) and then the grant proposal wants interesting science that's done
30 that way and high impact fo:r society and developing this (..) if you thought
31 you could still get the grant by: using your own internal metrics the:n you
32 probably submit it. If you thought that you wouldn't get the grant unless you
33 did something flashy a:nd quick and exciting you probably (.) you probably
34 think about a:m being a bit more relaxed on your own internal quality
35 metrics to try and get the grant. So maybe the sort of funding a:nd that
36 really is the driver. °if you've got as much money as you want I think° a lot of
37 scientists if they got a big grant they just do were happy to just carry on
38 doing what they want
39 I: mm
40 R: maybe quality is best the external measures of quality become less
41 important (.)°and the internal dominate more° yeah

The speaker not only argues that ideal science is best represented by internal metrics that each scientist has than by 'accepted scientific standards' (the metric model) which are 'in place' to represent it. He also implies that external metrics – impact in particular –encourage the reverse practices to the scientific rigour scientists would normally apply (lines 29-32). A third way is negotiated as the model in practice: a flexible use of internal and external metrics.

It is noteworthy that extract 5 comes from an interview with a participant who also has an administrative role in the faculty.

Extract 5 (Interview 6)

106 I: and what do you what do you think about these ((R just mentioned)) the
107 REF and the impact factor as metrics and assessment frameworks of quality?

R: e:m I think in some ways what I think about it >you know< it's the way we are going to be judged. And railing against the fact that it may not be >you know< I could have lots of debates about whether it's the right way of doing it but it's not going to make any difference to the way it's actually done
I: mm mhm
R: so I think it's a matter of making sure that for the university we fulfil those requirements but also for my own e:h I would always want to make sure I did what I saw as high quality research. So >for example< we have a list of journals that we should publish in. A:h I won't be using those journals simply because my research is in cancer >I'm in the department of electronic of electronic engineering< and consequently the top journals of electronic engineering include no cancer journals. Now if I were to publish my research in an electronic engineering journal they'd probably reject it. Because it's not in their remit. So that's where I would say I'd make the judgement because as far as I'm concerned the research has got to be in the right journal

Finally, extract 5 displays an even stronger case of quality being negotiated at the juncture of personal criteria and external measures.

In the preceding lines the speaker made reference to the 'impact factor' and the 'REF' in response to a question on how standards of quality in science are set. The interviewer picks this up. The response which then follows is formulated along the lines of opinion versus fact, the scientist's own view of high quality versus established standards (the metric model), (Initially, the speaker develops the argument that debating these standards will not change the way 'we' are going to be judged. This constructs these standards as debateable, the way 'we' are going to be judged as factual and positions the speaker as a powerless and a passive agent in this process, merely following these standards.

In lines 113-114, in the process of articulating how these standards are followed the speaker draws a distinction between those standards and her standards, as the 'for the university but also for my own' analogy is not made. So while 'for the university' 'we' make sure we fulfil those standards, (but) 'for her own' the speaker shifts footing to talk in the first person introducing (the distinction) her own view of 'what I saw as high quality science'. So while the position in the debate over the quality standards such as the 'impact factor' and the 'REF' is passive yet implying that the former are debateable, the practice of high quality research is presented as a matter of rational ('not in their remit', 'right journal') and calculated ('make the judgment') personal judgment. This also presents a distinction between the (external) ways of judgment and the (personal) right judgment. (The speaker: as a (better) judge of quality while the 'impact factor' and the 'REF' as mere requirements to be fulfilled). Most importantly, in so doing, the speaker attends to the norm of 'disinterestedness' (vs selfinterestedness)[6] as it is relevance ('the right journal') and not self-interested motivation and pursuit of wealth that guides the challenge of 'accepted scientific standards'. This registers a tension between ideal and metric models and suggests a third model – the practice model of science.

Overall, it could be argued that empirical evidence has highlighted that there is a tension between norms of ideal science, especially between quality (vs quantity) on the one hand and universalism (vs particularism), organised scepticism (vs organised dogmatism) and disinterestedness (vs selfinterestedness) on the other. This tension is grounded upon the application of normative ideals of science in established, institutional practices, and could therefore, classify under the normative binary of governance vs administration (Anderson et al 2007); in other words, the

---

[6] It should be mentioned though that marketing-related terms such as 'selling' or 'promoting' one's work were also used by participants, on the one hand as normative and directive by junior academics or PhD candidates and, on the other, as the conundrum of external quality assessment by more senior people.

misapplication of the norm of ‘governance’ of scientific practice, which renders it tedious and counter-productive ‘administration’ instead, but which, nevertheless, scientists are resisting through alternative action or rationalisation, as seen in the extracts above.

These findings are supported by a follow up study where academics and researchers in the social and natural sciences discussed activities related to collaboration and assessing publications in focus group sessions (N=4). The main patterns identified included first, a tension between the ideal and metric models, in particular as regards the norm of communality (vs secrecy) and the predicament of research output in the form of publications as the standard on which tenure and promotion in academia is based, encouraging self-interestedness. Namely, the more junior participants were in their academic career, the more cautious they were about sharing work. Mid career academics developed a line of disinterestedness ‘in principle’ and self-interest ‘in practice’; in other words, that collaborations advance science while also ensuring personal gain (in terms of advancement in one’s career or material returns such as employment and resources). Secondly, quality was juxtaposed to quantity with reference to citations. Namely, there was a tension with regards to the importance of the number of citations as a criterion of selecting work to read as well as a criterion of one’s reputation, while participants drew a clear distinction between citations and relevance in deciding the worth of journals (see also extract 5). Finally, participants commonly oriented to science as a divided enterprise. Namely, in cases of disagreement participants disclaimed their argument by claiming to speak for their discipline or research field.

5. Conclusion

In this paper we presented a study on the perceptions of scientists regarding quality in science. Using the conceptualizations of 'models of quality' in science, we have identified three models of science: metric, ideal, practice. We have looked at processes in place to assess quality in science and have identified a tension between the metric and ideal models. We then examined empirical data (one survey, interviews and focus groups) and focused on the model users – scientists – according to which there is a tension between the ideal and metric models and between scientists' understandings of quality in science and the metric model. It becomes apparent that this tension has an impact on the daily research lives of scientists. Whilst the theoretical associations with quality in science lean heavily towards the ideal science model, the interviews show very clearly that scientists are torn between this ideal of quality and the need for amassing publications for personal career advancement. Another clear tension exists between the openness of science and ensuring intellectual property rights. This was particularly important for younger scientists (see focus group results).

Overall, these contradictions and tensions indicate that different discourses are at work. Quality in interviews was negotiated at the juncture of internal, personal criteria for assessing quality and external, established measures. The internal criteria seem to correspond to Mertonian norms – the ideal model. The external measures seem to correspond to the metric model. This juncture indicates a tension or dilemma of expertise (Scientists versus University administration, REF, Journal Editors and Reviewers). Quality seems to be considered within this context and the negotiation constitutes a third way – the practice model.

One way to explain this is by making an analogy between quality and knowledge in science, drawing on the argument that knowledge is for use rather than just for contemplation, and actors – scientists in this case – have their own interests about how instruments work. So, instead of reaching closure in terms of what scientific practice is, actors can be seen as constantly seeking to extend culture in order to accommodate those interests, while interests themselves become the standards by which the products of scientific culture extensions are assessed (Pickering 1992, pp. 4-5). This argument seems to follow up to the latter point, whilst if it was indeed a straightforward case of actors' interests emerging as scientific standards, then tension would cease to exist.

Another way to explain this, thus, is through the binary of governance/administration, arguing that in the process of creating and establishing ways to assess quality, science metrics become a self-fulfilling prophecy, parts of the definition of quality. Yet, these parts do not develop organically as the tension indicates, and contingency measures are taken by scientists, e.g. open access approach to publications. The emergent model at work seems to be a compromise – the practice model – which ensures a rudimentary compliance with the metric model. This model is viable for producing quality according to scientists, and, thus, needs to be better supported by processes. Currently scientists are left to negotiate different conceptions of quality in science individually. Thus, while this model may emerge as the contingency strategy or 'third way' in academia, it cannot equally function as a solution to the tension observed for everyone involved, thus necessitating more widespread adoption of processes that support it.

6. References

Adler, R., Ewing, J. and Taylor, P. 2008. Citation Statistics. *A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*

Anderson et al. 2007. Normative Dissonance in Science. *Journal of Empirical Research on Human Research Ethics*: 3–14.

Banerjee, S. and Pedersen, T. 2003. *The Design, Implementation, and Use of the Ngram Statistics Package*. Proceedings of the Fourth International

Conference on Intelligent Text Processing and Computational Linguistics, February 17-21, 2003, Mexico City.

Boltanski, L. and Thevenot, L. 2006. *On Justification. Economies of Worth.* Princeton: Princeton University Press.

Charmaz, K. 1994. Discovering chronic illness: Using grounded theory. In B. Glaser (Ed.), More grounded theory methodology: A reader (pp. 65-93). Mill Valley, CA: Sociology Press.

Fanelli, D. 2009. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. PLoS ONE, 4(5): e5738.

Frey, B. S. 1994. How Intrinsic Motivation is Crowded out and in. *Rationality and Society*. Vol. 6 (3): 334-352.

Garfield, E. 1972. Citation Analysis as a tool in journal evaluation. Science, 178 (4060): 471-479.

Ghylin, K. M., Green, B. D., Drury, C. G., Chen, J., Schultz, J. L., Uggirala, A., Abraham, J. K., and Lawson, T. A. 2008. Clarifying the dimensions of four concepts of quality. *Theoretical Issues in Ergonomics Science*, 9, 1: 73–94.

Gilbert, G. N. and Mulkay, M. 1984. *Opening Pandora's Box: A Sociological Analysis of Scientist's Discourse.* Cambridge: Cambridge University Press.

Goldstein, H. and Spiegelhalter, D. J. 1996. League Tables and their Limitations: Statistical issues in comparisons of institutional performance. *Journal of Royal Statistical Society A*, 159 (3): 385-443.

Habermas, J. 2003. *Truth and Justification*. The MIT Press, Cambridge, Massachusetts.

Hirsch, J. E. 2006. An Index to Quantify an Individual's Scientific Researhc Output. Proc Natl Acad Sci USA, Vol. 102 (46): 16569-16573.

Kearns, J. N., & Fincham, F. D. 2004. A prototype analysis of forgiveness. Personality and Social Psychology Bulletin, 30: 838-855.

Latour, B. 2007. A Textbook Case Revisited—Knowledge as a Mode of Existence. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch and Judy Wajcman (Eds) *The Handbook of Science and Technology Studies*. (3rd Ed.) Cambridge, MA: The MIT Press.

Lofland, J. and Lofland, L. 1995. *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*, Belmont, CA: Wadsworth.

Merton, R. K. 1973. The Normative Structure of Science, in Merton, Robert K., *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago: University of Chicago Press.

Pickering, A. 1992. *Science as practice and culture*. Chicago: The University of Chicago Press.

Rayson, P. 2008. From key words to key semantic domains. International Journal of Corpus Linguistics, 13, (4): 519–549.

Rosch, E. 1978. Principles of Categorization. in Rosch, E. & Lloyd, B.B. (eds), *Cognition and Categorization*, Hillsdale: Lawrence Erlbaum Associates, Publishers: 27–48.

Sheldrick, G. M.2007. A Short Hisotry of SHELX. *Acta Crystallographica A*. 64(Pt.1): 112-122.

Squazzoni, F., Bravo, G. and Takacs, K. 2012. Does Incentive Provision Increase the Quality of Peer Review? *An Experimental Study. Research Policy*, in print.

Fanelli, D. 2010. Do pressures to publish increase scientists' bias? an empirical support from us states data. *PLoS ONE*, 5(4):e10271.

Fischbacher, U. and S. Gächter. 2010. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review* 100(1): 541-56.

Garfield, E. 1996. What is the primordial reference for the phrase 'publish or perish'? The Scientist, 10(12): 11.

Gray, A. and Bill Jenkins (2004) Government and administration: Too much checking, not enough doing? *Parliamentary Affairs*, 57(2): 269–287.

Hanfling, O. 1981. *Logical Positivism*. Columbia University Press.

Hodgson, H. 2004. A statement by The Royal Free and University College Medical School and The Royal Free Hampstead NHS Trust. Lancet, 363: 824

Mazlish, B. 1982. The Quality of 'The Quality of Science': An Evaluation. *Science, Technology, & Human Values*. 7: 42-52

Popper, K. 1934/1959. *The Logic of Scientific Discovery*. Routledge.

Smeyers, P. and N. C. Burbules. 2011. How to improve your impact factor: Questioning the quantification of academic quality. *Journal of Philosophy of Education*, 45(1).

Strauss, A., & Corbin, J. 1998. Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). Thousand Oaks, CA: Sage.

The Editors of The Lancet. 2 February 2010. Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, The Lancet, available at http://www.thelancet.com/journals/lancet/article/PIIS0140 6736(10)60175-4/fulltext.

Triggle, N. 24 May 2010. MMR doctor struck off register, British Broadcasting Channel, available at http://news.bbc.co.uk/2/hi/health/8700611.stm.

Wetherell, M. 1998. Positioning and Interpretative Repertoires: Conversation Analysis and Post-Structuralism in Dialogue. *Discourse & Society*, 9 (3): 387-412.